Evolution of molecular innovations in cyanobacterial light-perceiving systems

Dissertation

kumulativ

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

des Fachbereichs Chemie
der Philipps-Universität Marburg

vorgelegt von

Niklas Steube

aus Kassel

Marburg, 2023

Die vorliegende Dissertation mit dem Titel *Evolution of molecular innovations in cyanobacterial light-perceiving systems* wurde von November 2019 bis August 2023 am Max-Planck-Institut für Terrestrische Mikrobiologie in Marburg sowie am Fachbereich Chemie der Philipps-Universität Marburg in der Arbeitsgruppe *Evolutionäre Biochemie* unter der Leitung von Dr. Georg Hochberg angefertigt.

Vom Fachbereich Chemie der Philipps-Universität Marburg
(Hochschulkennziffer 1180) als Dissertation angenommen am: 01.11.2023

Erstgutachter:	Dr. Georg Hochberg
Zweitgutachter:	Prof. Dr. Peter Graumann

Tag der Disputation: 02.11.2023

Erklärung zur Promotion

Hiermit versichere ich, Niklas Steube, dass ich die vorliegende Dissertation mit dem Titel *Evolution of molecular innovations in cyanobacterial light-perceiving systems* selbstständig, ohne unerlaubte Hilfe Dritter, angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht. Dritte waren an der inhaltlich-materiellen Erstellung der Dissertation nicht beteiligt (mit Ausnahme der genannten Autoren in den Originalpublikationen); insbesondere habe ich hierfür nicht die Hilfe eines Promotionsberaters in Anspruch genommen. Kein Teil dieser Arbeit ist in einem anderen Promotions- oder Habilitationsverfahren verwendet worden. Mit dem Einsatz von Software zur Erkennung von Plagiaten bin ich einverstanden.

Marburg, 31.08.2023

Niklas Steube

Erklärung zur kumulativen Dissertation

Wir, Niklas Steube (Doktorand) und Dr. Georg Hochberg (Betreuer), versichern hiermit, dass die im kumulativen Teil der Dissertation mit dem Titel *Evolution of molecular innovations in cyanobacterial light-perceiving systems* aufgeführten Anteile der Autoren an den verfassten Publikationen und Manuskripten korrekt und vollständig dargelegt sind.

Der kumulative Teil der Arbeit umfasst: Chapter II **Original Research Publications**

Marburg, 31.08.2023

Niklas Steube Dr. Georg Hochberg

This page intentionally left blank

Content

Chapter I **Introduction**

Chapter II **Original Research Publications**

Evidence for an early green/red photocycle that precedes the diversification of GAF domain photoreceptor cyanobacteriochromes

Results

Discussion

Fortuitously compatible protein surfaces primed allosteric control in cyanobacterial photoprotection

Results

Chapter III **Discussion**

Appendix

Zusammenfassung

Evolution molekularer Innovationen in cyanobakteriellen Systemen der Lichtwahrnehmung

Neue funktionale Eigenschaften sind evolutionär von besonderer Bedeutung und doch paradox: Wie kann Evolution etwas Innovatives schaffen, wenn sie nur mit Variationen von etablierter Biologie arbeiten kann? Neofunktionalisierung von Proteinen nach vorausgegangener Genduplikation ist eine gängige Erklärung, angetrieben durch natürliche Selektion für eine neue, potenziell innovative Funktion. Es ist jedoch fraglich, ob wegweisende Neuheiten tatsächlich durch bloße adaptive Diversifizierung bestehender Proteine erklärt werden können. In dieser Arbeit haben wir das Paradox der molekularen Innovation mit Hilfe molekularer Phylogenetik in zwei Originalveröffentlichungen untersucht.

Der erste Artikel erforschte die Evolution von Cyanobakteriochromen (CBCRs), einer Klasse von Phytochromen, die ausschließlich in Cyanobakterien vorkommen. CBCRs haben die innovative Fähigkeit erlangt, kollektiv das gesamte Spektrum des sichtbaren Lichts mit einem Einzeldomänen-Protein wahrzunehmen, im Gegensatz zu den kanonischen Phytochromen mit drei Domänen, die in erster Linie auf rote und fern-rote Lichtsignale reagieren. Mit Hilfe von Ahnensequenzrekonstruktion (ASR) und biochemischer Verifizierung der wiedererweckten Proteine haben wir gezeigt, dass der letzte gemeinsame Vorfahre der CBCRs reversibel auf Grün- und Rotlicht-Signale reagierte. Latente Blaulicht-Wahrnehmung sowie die Fähigkeit, alternative Chromophore zu binden, gepaart mit der minimalistischen Domänenarchitektur, könnten die gewaltige Diversifizierung der CBCRs ermöglicht haben. Dies deutet darauf hin, dass molekulare Innovationen möglicherweise durch eine Verringerung von Proteinkomplexität erreicht werden können, wodurch sich wiederum neue Wege im Sequenzraum für neue Funktionen, wie beispielsweise breitere Farbwahrnehmung, auftun.

Der zweite Artikel befasste sich mit der Entwicklung einer neuartigen allosterischen Regulierung beim Lichtschutz von Cyanobakterien durch direkte Protein-Protein-Interaktion. Es ist unklar, ob die dafür erforderlichen Protein-Oberflächenkompatibilitäten nur durch Selektion in kleinen Schritten oder auch zufällig entstehen können. Hier haben wir ASR und biophysikalische Proteincharakterisierung genutzt, um die Entwicklung der allosterischen Interaktion zwischen dem orangefarbenen Carotinoid-bindendem Protein (OCP) und seinem nicht verwandten Regulator, dem Fluoreszenzrückgewinnungsprotein (FRP), zu rekapitulieren. Diese Interaktion entwickelte sich, als ein Vorläufer von FRP horizontal von Cyanobakterien erworben wurde. Die Vorläufer von FRP konnten bereits mit OCP interagieren und es regulieren, noch bevor diese Proteine in einem Ur-Cyanobakterium erstmals aufeinandertrafen. Die OCP-FRP-Interaktion nutzt dabei eine uralte Dimer-Schnittstelle in OCP, die auch schon vor der Aufnahme von FRP in das Lichtschutzsystem bestand. Dies zeigt, wie einfach Evolution komplexe regulatorische Systeme aus bereits existierenden Komponenten aufbauen kann, selbst ohne vorausgegangene Genduplikation.

Zusammenfassend haben wir gezeigt, dass Zufallsereignisse in der Proteinevolution eine unterschätzte Rolle spielen können und tatsächlich zu wegweisenden biologischen Innovationen führen.

Abstract

Novel functional features are prominent throughout evolution, yet paradoxical: how does evolution create something innovative when all it can work with is variation of established biology? Neo-functionalization of proteins after gene duplication is one common explanation, driven by natural selection for a new, potentially innovative function. However, groundbreaking novelty may not be explained by adaptive diversification of existing proteins. In this thesis, we tackled the paradox of molecular innovation with molecular phylogenetics in two original research publications.

The first article examined the evolution of cyanobacteriochromes (CBCRs), a class of phytochromes found exclusively in cyanobacteria. CBCRs gained the innovative ability to collectively sense the entire spectrum of visible light with a single-domain protein, in contrast to canonical tri-domain phytochromes that respond primarily to red- and far-red light signals. Using ancestral sequence reconstruction (ASR) and biochemical verification of resurrected proteins, we showed that the last common ancestor of CBCRs responded reversibly to green- and red-light signals. Latent blue-light perception and the ability to bind alternative chromophores, coupled with the minimalistic domain architecture may have enabled the vast diversification of CBCRs. This indicates that molecular innovation can potentially be achieved by reducing protein complexity, which may open up sequence space for new functions, such as broader color perception.

The second article focused on the evolution of a novel allosteric regulation in cyanobacterial photoprotection by direct protein-protein interaction. It is unclear whether such required protein surface compatibilities can only be built by selection in small incremental steps, or whether they can also emerge fortuitously. Here, we used ASR and biophysical protein characterization to retrace the evolution of the allosteric interaction between the orange carotenoid protein (OCP) and its unrelated regulator, the fluorescence recovery protein (FRP). This interaction evolved when a precursor of FRP was horizontally acquired by cyanobacteria. FRP’s precursors could already interact with and regulate OCP even before these proteins first encountered each other in an ancestral cyanobacterium. The OCP–FRP interaction exploits an ancient dimer interface in OCP, which also predates the recruitment of FRP into the photoprotection system. This shows how evolution can easily fashion complex regulatory systems from pre-existing components, even without prior gene duplication.

Together, we have shown that chance events may play an underestimated role in protein evolution and can indeed lead to groundbreaking innovations in biology.

Chapter I

Introduction

1| Molecular innovation of novel functional features

Novel functional features are a prominent key phenomenon in the diversification of life. Molecular innovations like the ability of a photoreceptor to perceive new wavelengths or the implementation of novel regulatory control over existing biological systems allow organisms to evolve by executing ultimately more sophisticated tasks. Such new capabilities may be important to cope with changing ecological conditions or to outperform competitors in shared habitats[1]. However, the origin of novelty typically represents a paradox[2]: how can something new and innovative appear when evolution can only work with something old and proven (Fig. 1a,b)?

Novel phenotypic features like vison or limbs have emerged independently several times during evolution[3]. Light-sensing organs always perform a similar function, but are physiologically and morphologically highly diverse[4]. However, their development is mostly controlled by the same set of highly conserved transcription factors[5]. The same is true for limb development[6]. These *deep homologies* explain parallel evolution of innovative features by using pre-existing regulatory protein circuits. In contrast, how proteins evolve new functions is mostly unknown.

A common explanation for novel protein features is the exaptation-amplification-diversification (EAD) model[2,7,8]: proteins often perform moonlighting functions that are not their essential task, but may appear as a by-product of some degree of promiscuity[9,10]. By co-opting such a side reaction (exaptation) and amplification of the corresponding gene under changing ecological conditions, random mutations can occur in the copied version of the gene without affecting the essential protein's primary function still encoded in the original version of the gene still present. The new homologous versions of the protein could then be selected for increased efficiency of the moonlighting task in several rounds of accumulating mutations to eventually neo-functionalize as a new protein with a novel, potentially innovative function (diversification).

The EAD model could in principle explain the appearance of potentially simple transitions like the photoreceptors that sense new colors, because such changes may be accomplished through only a small number of mutations that would each embody a phenotype strong enough to be selected for. Real novelties, like a new regulatory protein that has to perform a completely new task, may in turn be too complicated and would necessitate intermediate mutational steps that may not be tolerated by purifying selection.

Further, it seems that the multi-step EAD model could in fact hamper real groundbreaking innovation, because every mutational step has to build on the former and every intermediate along the trajectory towards a new function has to provide at least some functional benefit to be selected for[11]. But how can something truly novel appear in biology that is not just a variation of something that already exists? Where do new regulatory components come from and how can they be integrated into already existing and fully functional biological systems? Finally, is natural selection actually the main driver of biological innovation?

Directed protein evolution experiments often exploit the EAD paradigm by engineering a known protein's moonlighting function into the desired main function[12]. To achieve this, they usually set up an (unnaturally) strong selection pressure for that one specific trait that should be improved and run iterative mutational cycles till the desired functionality is reached[13–17]. Such experiments have shown that great improvements and major functional transitions can be achieved quite rapidly in only a few mutational steps[13–17]. However, they tell little about the actual evolution of proteins under natural conditions where more than one controlled strong selection pressure shapes the evolution of the protein. But raised with this kind of targeted selection-driven experiments, biochemists tend to find adaptive explanations for biological improvements or new functional features. Consequently, natural selection may potentially be overestimated as the main driving force of evolutionary innovation.

Novel protein-protein interactions (as mediators of innovative features) can get entrenched quite fast[18–21]: once a novel interaction is established, hydrophobic mutations can occur in the interface that were former not tolerated because the participating residues would have been directly exposed to the aqueous environment before. Once substituted, the novel interaction is entrenched, meaning that both proteins are henceforward dependent on the interaction because they are not stable any longer without their new partner. Such novel complexity may also occur neutrally without any functional improvement, for no first-order adaptive reason[18–20]. Further, molecular complexity increased in eukaryotic V-ATPases and hemoglobin in short genetic trajectories of only one and two historical substitutions, respectively[22,23].

Taken together, this shows that evolutionary routes to molecular novelty can be short and suggests that chance events may also play an important role in natural protein evolution. Can molecular innovation also happen through *happy accidents*[24]? Besides, are more sophisticated systems always more complex?

To understand causation and mechanisms of functional, molecular innovations that appear on reasonable evolutionary timescales of hundreds of millions of years without a single controlled selection pressure, we need to study the characteristics of old enzymes and their transitions into extant ones, the molecular foundation for new biological features. But where could we possibly get them from?

Although paleoproteomics (the study of ancient proteins from fossil record) is an innovative and rapidly growing field, fossilized proteins are scarce and fossils containing DNA (the blueprint for proteins) are limited[25]. We are further not (yet) able to resurrect entire extinct species to bring back their old biology, although people are actively trying[26].

However, we have molecular phylogenetics. We can order biological relations not only on the organismal level (like classical phylogenetics), but also on the molecular, the protein level[27]. With such inferred protein phylogenies, we can resurrect old enzymes of long extinct species with the help of a phylogenetic method called ancestral sequence reconstruction (ASR)[28]. This allows to study protein function transitions and the emergence of molecular innovations through evolutionary time and can help dissecting the driving forces behind their appearances.

In this thesis, we investigated molecular innovations in two cyanobacterial light-perceiving systems that are remarkable, but still simple enough to study experimentally: novel multi-color sensing in cyanobacteriochromes (CBCRs), and the evolution of a new regulatory protein (FRP) in cyanobacterial light protection. We later discuss, if the EAD model can help explaining the evolution of these innovations, or if we need to consider new explanatory approaches to understand the evolution of biological novelty.

2| Maximum likelihood protein phylogenetics

Phylogenetics is the study of relationships between entities that in biology reflect the evolutionary links between species. By comparing morphological or physiological character traits of organisms, biologists aim to understand relatedness between species to draw conclusions about their evolution. Analyzing the presence or absence of certain traits and grouping species with similar ones (synapomorphies) together, allow to order the living world and to draw conclusions about their origin[29].

To represent evolutionary history, biologists since Darwin in 1837 draw phylogenetic trees that start at a common point, called the root[30]. From this root, the tree bifurcates every time a new group of organisms, that share one specific trait, evolves. The emerging branches bifurcate in the same manner till every species sits on its own branch and a tree-like structure develops with leaves (tips) representing extant species connected by branches that unite repeatedly at junctions (nodes) down to the root[31].

A phylogenetic tree should generally be constructed in a way that represents the least amount of character trait changes necessary (rule of maximum parsimony)[32]. This rule bases on a universal parsimony principle also known as Ockham's razor that leads back to the medieval philosopher William of Ockham and is mostly accepted as a basic explanatory principle in all sciences[33].

Each node on a rooted phylogenetic tree represents the last common ancestor (LCA) of the derived groups that together share the one trait that unites all its descendants. Such a group shares common ancestry and is called monophyletic. Species in monophyletic groups are always more closely related to each other than to any other species on the tree (Fig. 1c)[31].

Simple phenotypic trees as described above come with some drawbacks, e.g. it is not evident which traits to choose to analyze as complex organisms may have thousands of comparable traits. Further, most traits have only limited states like simple present or absent distinctions and some are even misleading, as traits can also evolve several times independently without common descent (homoplasy). Besides, there is only a direction, but no time information on such trees (branch lengths are arbitrary), and it is often not self-evident where to place the root, the starting point of evolution.

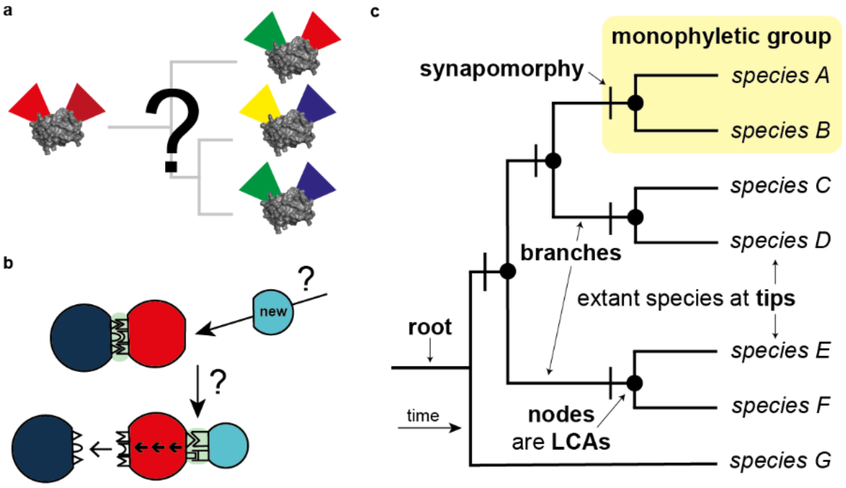


Fig 1| Investigation of novel functional features with molecular phylogenetics. **a**,**b**, Molecular innovations like multi-color perception in a dual-color sensing photo-receptor (**a**) or novel allosteric regulation (**b**) appear paradoxical, as something new seemingly evolves out of something old. Protein structures used in **a** for illustration only is 4GLQ (PDB ID)[34]. **c**, Example phenotypic phylogenetic species tree. LCAs, last common ancestors.

The possibility of DNA sequencing revolutionized phylogenetics. Instead of phenotypic traits, nucleic or amino acid sequences can be used to compare species[27]. Genes on DNA code for proteins that characterize all living beings precisely on the molecular level. Proteins are composed of a single chain of dozens or up to thousands of amino acids. Each amino acid has a certain position in the protein and thus represents a single, distinguishable molecular trait of that one protein. At each position, there are 20 possible states, corresponding to the 20 canonical amino acids (aa).

For an example protein of 267 aa (the median length of a bacterial protein[35]), there are 20^{267} combinatorial versions of that one protein. In addition, an organism typically features hundreds to thousands of different proteins. This massive amount of comparable data for each single (sequenced) species allows to overcome the drawbacks of phenotypic trees: for a single protein, each of its hundreds of molecular traits (represented by one aa position in the protein) has 20 possible states that may vary between species. All traits can be analyzed in parallel (without choosing any) and they feature the exact same possible 20 states. This allows to analyze substitution rates (state changes between homologs) and thus adds a temporal dimension (scalable branch lengths). Homoplasy is further rarer on the protein sequence level, compared to phenotypic traits[20].

To unravel the evolution and diversification of species, we may consider the information of all their proteins (or at least of the ones they share). However, we can also look at distinct evolutionary histories of single proteins which may not be identical to the species' history and could thus give further insides into how evolution works on the molecular level, the level of individual proteins. The analysis of a certain protein's evolutionary history is performed in three main steps (Box 1)[36].

1) Choice of Protein and Taxonomic Sampling Range of Interest

Every analysis starts with a protein of interest and its taxonomic range, meaning in which taxonomic groups the protein's evolution should be analyzed. The protein has to meet the minimal requirements (certain conserved length; presence in most of the species, but also decent sequence divergence within the taxonomic sampling range; adequate sequence availability). There are no hard rules on these requirements and they usually need to be judged on a by-case basis, depending on the precise evolutionary question one seeks to answer[36].

2) Sequence Data Gathering

The aa sequence of the focal protein of interest serves as a query to find proteins with similar sequences in related species (orthologs) by using local alignment search tools against sequence data bases like BLAST[37]. Homolog hits (alongside the query) get aligned with software like MUSCLE, meaning that the sequences of different homologs are arranged in a way that homologous aa blocks between sequences are listed one below the other without changing the order within one sequence by introducing gaps[38]. The result is a multiple sequence alignment (MSA). Dense taxon sampling is crucial, meaning that all major taxonomic groups within the sampling range are covered and no group is overrepresented. This is secured by comparison with a known species phylogeny and including sequences from every major taxonomic group of interest.

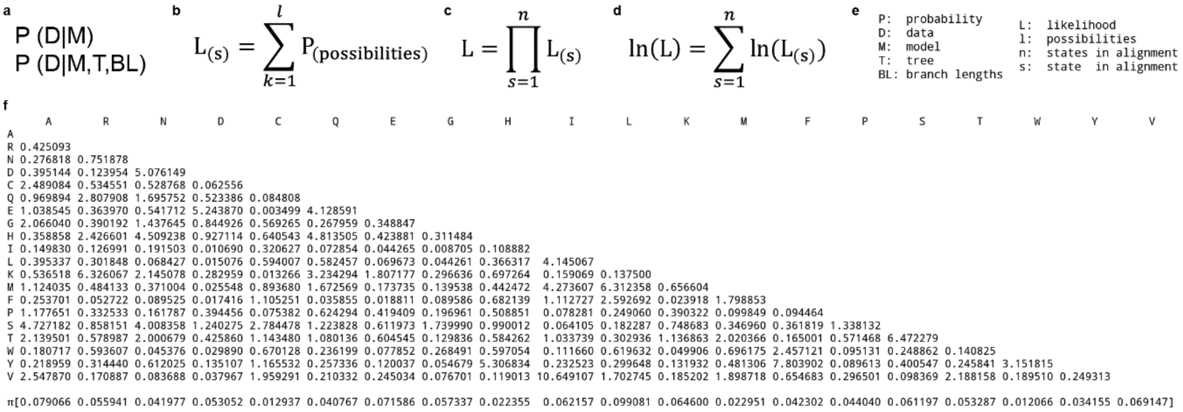
3) MSA Trimming

Gaps in the MSA are treated as missing data in the following analyses and should be minimized by deleting whole sequences with anomalous length and stretches of gaps (deletions) or of sequence (insertions) that are linage-specific, meaning that they only appear in a certain small monophyletic group (linage) of species. The trimmed MSA is used to computationally infer the phylogenetic protein tree.

Box 1| Analysis of a protein's evolutionary history with molecular phylogenetics.

In 1973, Joseph Felsenstein published the idea to use maximum likelihood estimation to infer phylogenetic trees[39]. Maximum likelihood (ML) is a statistical method of estimating the parameters of an assumed probability distribution, given some observed data and a statistical model. This is achieved by altering the free parameters in a way that maximizes the corresponding likelihood function. In case of a protein tree, the observed data is the trimmed MSA and the parameters to be fitted are the tree topology (the branching pattern) and the lengths of the branches (the average substitution rates)[39]. The underlying statistical model is a model of how proteins evolve, representing the mechanism of molecular change. Instead of phenotypic traits, it is assumed that molecular sequence changes mostly at random[27]. But the abundance of different aa and the possibilities that certain aa change to specific other ones differ significantly, depicting the basis for the statistical models of protein evolution.

These protein models feature a composition part, that is how frequent certain aa appear in a protein as well as a process part, that is how frequent aa change from one to another. The composition part is the sum of all frequencies for each single aa that always add up to 1. The process part is a matrix of 20 x 20 rate values (corresponding to the 20 canonical aa). There are several evolutionary models available that have been empirically derived from different data sets (Fig. 2a-f)[39].



a

P (D|M)
P (D|M,T,BL)

b

$$\mathrm{L}_{(\mathrm{s})} = \sum_{k=1}^{l} \mathrm{P}_{(\mathrm{possibilities})}$$

c

$$\mathrm{L} = \prod_{s=1}^{n} \mathrm{L}_{(\mathrm{s})}$$

d

$$\ln(\mathrm{L}) = \sum_{s=1}^{n} \ln(\mathrm{L}_{(\mathrm{s})})$$

e

P: probability
D: data
M: model
T: tree
BL: branch lengths

L: likelihood
l: possibilities
n: states in alignment
s: state in alignment

f

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A																				
R	0.425093																			
N	0.276818	0.751878																		
D	0.395144	0.123954	5.076149																	
C	2.489084	0.534551	0.528768	0.062556																
Q	0.969894	2.807908	1.695752	0.523386	0.084808															
E	1.038545	0.363970	0.541712	5.243870	0.003499	4.128591														
G	2.066040	0.390192	1.437645	0.844926	0.569265	0.267959	0.348847													
H	0.358858	2.426601	4.509238	0.927114	0.640543	4.813505	0.423881	0.311484												
I	0.149830	0.126991	0.191503	0.010690	0.320627	0.072854	0.044265	0.008705	0.108882											
L	0.395337	0.301848	0.068427	0.015076	0.594007	0.582457	0.069673	0.044261	0.366317	4.145067										
K	0.536518	6.326067	2.145078	0.282959	0.013266	3.234294	1.807177	0.296636	0.697264	0.159069	0.137500									
M	1.124035	0.484133	0.371004	0.025548	0.893680	1.672569	0.173735	0.139538	0.442472	4.273607	6.312358	0.656604								
F	0.253701	0.052722	0.089525	0.017416	1.105251	0.035855	0.018811	0.089586	0.682139	1.112727	2.592692	0.023918	1.798853							
P	1.177651	0.332533	0.161787	0.394456	0.075382	0.624294	0.419409	0.196961	0.508851	0.078281	0.249060	0.390322	0.099849	0.094464						
S	4.727182	0.858151	4.008358	1.240275	2.784478	1.223828	0.611973	1.739990	0.990012	0.064105	0.182287	0.748683	0.346960	0.361819	1.338132					
T	2.139501	0.578987	2.000679	0.425860	1.143480	1.080136	0.604545	0.129836	0.584262	1.033739	0.302936	1.136863	2.020366	0.165001	0.571468	6.472279				
W	0.180717	0.593607	0.045376	0.029890	0.670128	0.236199	0.077852	0.268491	0.597054	0.111660	0.619632	0.049906	0.696175	2.457121	0.095131	0.248862	0.140825			
Y	0.218959	0.314440	0.612025	0.135107	1.165532	0.257336	0.120037	0.054679	5.306834	0.232523	0.299648	0.131932	0.481306	7.803902	0.089613	0.400547	0.245841	3.151815		
V	2.547870	0.170887	0.083688	0.037967	1.959291	0.210332	0.245034	0.076701	0.119013	10.649107	1.702745	0.185202	1.898718	0.654683	0.296501	0.098369	2.188158	0.189510	0.249313	
π[	0.079066	0.055941	0.041977	0.053052	0.012937	0.040767	0.071586	0.057337	0.022355	0.062157	0.099081	0.064600	0.022951	0.042302	0.044040	0.061197	0.053287	0.012066	0.034155	0.069147]

Fig. 2| ML estimation to infer phylogenetic trees. a, ML estimates the probability (P) of the data (D), given the model (M). For protein trees, D is the MSA; M is a model of protein evolution and the tree topology (T) with scalable branch lengths (BL). **b**, The likelihood function for every state in the MSA is summed over all possible substitutions on one tree. **c**, The overall tree likelihood is the product of individual likelihoods of all states in the MSA. **d**, To prevent arithmetic underflow during computational processing, likelihoods are transformed into log values, and summed up. **e**, Variable definitions. **f**, Symmetrical substitution rate matrix (in log-odds) and aa frequencies (π) of the LG model of protein evolution that was empirically derived from 3,912 MSAs with more than 50,000 protein sequences[40]. Single letter aa code was used (A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; Q, glutamine; E, glutamic acid; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine).

To infer a protein tree, a phylogenetic software fed with a trimmed MSA initially creates an unrooted tree by neighbor joining (NJ). NJ is a fast method that is not computationally demanding and clusters the sequences in the MSA by a distance matrix[41]. Next, a ML algorithm fits the free tree parameters (tree topology and branch lengths) in iterative steps to infer a tree with the highest likelihood, given the model of protein evolution. To find the best-fit model, the software first tests different implemented empirical models using the Akaike information criterion (AIC) that (in its easiest interpretation) provides a statistical measure of fit between the model and the data while penalizing for over-parameterization and correcting for small sample sizes[42].

The computational demand for ML algorithms is high. It is impossible to iterate over all possible trees in decent time. To still maximize the free parameters, heuristic hill-climbing approaches are used that start with the NJ tree and tolerate parameter changes only if they increase the likelihood of the new tree until maxima for all free parameters are reached and the tree likelihood cannot be improved any more. Hill-climbing approaches are prone to get trapped in local maxima, meaning that found maxima are only valid for the particular starting tree, but (higher) global maxima may exist that cannot be reached from that particular starting point. To minimize this problem, we use subtree pruning and re-grafting (SPR) moves during the hill-climbing process. This means that random branches get removed from the initial tree and transplanted to a different position of the remaining tree. Only if such rearrangements improve the tree likelihood, the changes are accepted. The SPR moves help best to overcome local maxima, although they still cannot totally rule them out[43].

Another common difficulty for inferring molecular phylogenies is heterogeneity of evolutionary (substitution) rates among sites in the MSA, meaning that specific sites of the protein evolve slower than others[42]. The most prominent example is the start codon of a gene that is under strong purifying selection, as it is essential to initiate translation. The first state in a protein is thus (almost) always methionine with an extremely slow evolution rate of typically 0. Other examples for slow evolving sites are the active site of a protein or other conserved structural features that are essential for function. Mostly unstructured regions like loops, linkers, or terminal extensions in turn typically show faster rates of evolution. This heterogeneity is critical, because evolutionary change from one protein to another is displayed by a single branch length on the phylogenetic tree.

To accurately correct for rate heterogeneity among sites, we would need an evolutionary rate parameter for every position in the MSA. As this would over-parameterize our model and would be computationally highly demanding, we use a gamma distribution of rates instead. First, the frequencies of evolutionary rates over all states are calculated and grouped into (typically four) distinct categories. This allows to model a probability distribution (the gamma distribution) of sub-rates among sites that is characterized by a single parameter. With this alpha parameter evolutionary rate heterogeneity among sites can be sufficiently accounted for with the addition of a single parameter in the model of protein evolution[42].

These adjustments described above help to improve the tree inference, but ML is still a statistical method that infers the most likely phylogenetic history for our protein of interest. However, this may not be the exact historically accurate protein history. We have to keep in mind that we reconstruct evolutionary history of typically over thousands of millions of years using only sequence information that is conserved in proteins of species that are alive today. To account for these historical uncertainties, we test our tree inferences with statistical methods and display confidence values for the inferred tree topology by using the bootstrap approach[44]. Bootstrapping is a statistical re-sampling method that allows to test for accuracy of our data by random sampling with replacement. For a protein phylogeny, we test if our data set (the MSA) is representative of the underlying (unknown) population from which it was drawn, or if the tree topology relies on only a few specific sites in the MSA. This is achieved by computationally generating pseudo-MSAs that contain the same amount of sequences of unaltered length, but with randomly sampled (with replacement) character states (columns in the MSA), and hence inferring 100 bootstrap trees.

Felsenstein Bootstrap Probabilities (FBPs) for every node on the initial ML tree topology are calculated by counting up the amount of bootstrap trees that feature that specific node with a value from 100 (indicating that all bootstrap trees agree on the node) to 0 (indicating that no bootstrap replicate recapitulated it)[44]. As this method cannot distinguish between major topological discrepancies or minor branch rearrangements near the node, we further infer Transfer Bootstrap Expectations (TBEs). TBEs represent the node support as a percentaged value by taking into account the extant of topological rearrangements at specific nodes between the initial ML and the bootstrap trees[45]. As the TBEs tend to euphemize bootstrap support if one side after a split at a node is huge, we additionally test branch support with approximate likelihood ratio tests (aLRT). These statistics compare the likelihood of the initial branch arrangement with the second most likely arrangement around that node and give a measure of importance for a specific node on the tree[46].

In the end we get the most likely phylogenetic protein tree with three statistical support values. However, this tree is still unrooted, meaning that the evolutionary history is lacking a direction. The most common method to root a phylogenetic tree is outgroup rooting[47]. This means, initially adding aa sequences of homologs to the MSA that belong to species that are more distantly related from all species within the taxonomic sampling range. Their homologs should have diverged from the focal proteins within the taxonomic sampling range earlier than the existence of their LCA, and thus provide a time point that is evolutionary older than any of the focal proteins of interest and suits as an evolutionary starting point of the tree[47]. By placing the root between this outgroup and the remaining sequences, a direction is set that allows to draw conclusions about the evolution of the studied protein.

Discrepancies between such protein trees and a known species phylogeny help to identify evolutionary events in the protein's history: a gene duplication also duplicates the corresponding species topology after the duplication node on the protein tree whereas a gene loss in certain organismal groups prune proteins of those species from expected branches on the protein tree. In addition, horizontal gene transfers can be identified, if a protein sequence nests within or is sister to a group of proteins of distantly related species. This allows to unravel the individual evolutionary histories of specific proteins and can help to identify the driving forces behind evolutionary events (Fig. 3a,b).

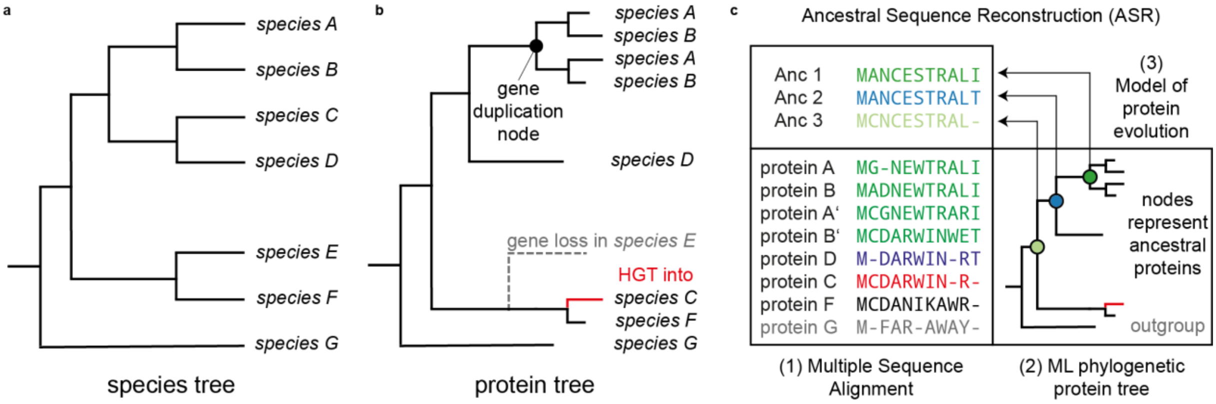


Fig. 3| Protein trees reveal evolutionary events and enable the resurrection of ancestral proteins. a, Example species tree (Fig. 1c). **b**, Comparison with a corresponding tree for one specific protein reveals gene duplications, gene losses, and horizontal gene transfer (HGT) events in the protein's evolutionary history. The outgroup *species G* that is distantly related to all other species on tree **a** also roots the protein tree. **c**, Reconstruction of ancestral aa sequences at internal nodes on the protein tree by ASR to resurrect and characterize ancestral proteins in the laboratory.

3| Resurrection of ancestral proteins and their biochemical characterization

To finally resurrect ancestral proteins to biochemically characterize them in the laboratory, we already have all necessary input from the ML tree inference: a phylogenetic protein tree, the underlying MSA, and the best-fit model of protein evolution. The only other crucial information needed is the exact evolutionary time point to which the ancestral protein should date back. We are generally interested in LCAs of certain protein groups that existed prior to evolutionary events like gene duplications or horizontal gene transfers. These LCA proteins are represented by internal nodes on the protein tree and their evolutionary distances are exactly defined (in average substitutions per site) by the branch lengths connecting them to the neighboring nodes or tips. We can reconstruct every tree node that has at least one preceding node and two descending nodes or tips. This is true for all internal nodes on the tree, except for the root node (that is arbitrarily positioned on the root branch, conventionally in the middle). A ML algorithm infers the posterior probabilities for each state at every position in the LCA proteins. By taking the aa with the highest posterior probability at every position, we get the most likely aa sequence for every ancestral protein at every internal node on the protein tree (Fig. 3c)[28].

To then resurrect an ancestral protein, we back-translate its aa sequence into a codon-optimized nucleotide sequence for *Escherichia coli*. The DNA sequence is cloned into an expression vector with an 6x histidine epitope tag and transformed into *E. coli*. The protein is over-produced and purified by affinity purification and size exclusion chromatography. With the purified protein in solution, biochemical assays of any kind can be performed, like with any extant protein. In our case, we mainly characterized the ancestral proteins in terms of their behavior upon light irradiation with UV-Vis spectroscopy that measures light absorbance in the ultra-violet to red wavelength spectrum of light.

As the ancestral protein sequences are statistical estimations, we further characterize alternative ancestors that feature the aa state with the second highest posterior probability at ambiguous sites. By comparing their properties with the initial ancestral proteins, we test the robustness to statistical uncertainties in the reconstructions, analog to bootstrapping the tree[48].

4| Cyanobacterial light-sensing proteins as model systems to study innovation

Cyanobacteria are photo-autotrophic, gram-negative bacteria that evolved oxygenic photosynthesis, a mechanism to convert light into chemical energy by using carbon dioxide and water molecules to produce energy-rich carbohydrates and releasing oxygen into the atmosphere[49]. With this ability to use sunlight as a food source came the necessity to anticipate light. The mostly aquatic cyanobacteria have to move towards light sources in the water column, but must also protect themselves from excessive irradiation that causes photo-damage.

Photoreceptor proteins sense incident light and trigger downstream signal transduction events in photo-active species like cyanobacteria[50]. Phytochromes are a superfamily of photoreceptors that bind a linear bilin molecule as a chromophore that reversibly interconverts between two isoforms. This allows to sense two distinct wavelength of the incident light, mostly red and far-red. Phytochromes are found in plants, fungi and bacteria, and show a typical tri-domain architecture[51,52]. However, minimal versions of these bilin-bound phytochromes that only require a single domain, but collectively sense the whole spectrum of visible light have exclusively evolved in cyanobacteria[53]. The expansion of the light perception spectrum of these cyanobacteriochromes (CBCRs) represent a remarkable molecular innovation. In the first original publication of this thesis, we investigated the evolution of CBCR proteins. We asked how the light perception of the LCA of all CBCRs differed from canonical phytochromes and sought to elucidate how they diversified into sensing the whole color palette with only a single functional domain.

Photoprotection in cyanobacteria is mediated by the orange carotenoid protein (OCP)[54]. High light activates OCP by causing a conformational change in the protein[55]. Only when activated, OCP binds to the light-harvesting antenna complexes to dissipate excess energy as heat[55,56]. OCP's recovery into the resting state is a passive progress in most OCP paralogs[55,57], but one of them (OCP1) relies on an allosteric regulator for back-transformation[58,59]: the fluorescence recovery protein (FRP) terminates the interaction with the antenna complex and strongly accelerates photo-recovery of OCP1[58,60]. This novel allosteric control via direct protein-protein interaction provides an innovative new functional feature in cyanobacterial photoprotection. In the second publication, we studied how and when in cyanobacterial history this new interaction between these two initially unrelated proteins evolved.

5| Aims and structure of this thesis

Novel functional features are common, but seemingly paradoxical at the same time. This thesis features two recent publications that examined the origins and the evolution of two molecular innovations in cyanobacterial light-sensing systems and aims to first recapitulate common explanations for functional innovation in biology and to explain the authors' approach to the paradox via molecular phylogenetics. Second, to test if these explanations hold true for the two model systems studied, and finally, to discuss new perspectives on the origin of innovative functional features in biological systems.

This cumulative thesis is structured into three chapters: Chapter I provides introductory explanations about the rationale behind the two published studies in Chapter II, and concludes with a final discussion in Chapter III. Figures, Tables, and references are numbered separately for each chapter or article. References can be found at the end of Chapter I and III or at the end of each article in Chapter II.

6| References

1. Wagner, A. The molecular origins of evolutionary innovations. *Trends in Genetics* **27,** 397–410 (2011).
2. Kassen, R. Experimental evolution of innovation and novelty. *Trends in Ecology & Evolution* **34,** 712–722 (2019).
3. Shubin, N., Tabin, C. & Carroll, S. Deep homology and the origins of evolutionary novelty. *Nature* **457,** 818–823 (2009).
4. Schwab, I.R. The evolution of eyes: major steps. The Keeler lecture 2017: centenary of Keeler Ltd. *Eye* **32,** 302–313 (2018).
5. Gehring, W.J. New perspectives on eye development and the evolution of eyes and photoreceptors. *Journal of Heredity* **96,** 171–184 (2005).
6. Shubin, N., Tabin, C. & Carroll, S. Fossils, genes and the evolution of animal limbs. *Nature* **388,** 639–648 (1997).
7. Francino, M.P. An adaptive radiation model for the origin of new gene functions. *Nature Genetics* **37,** 573–577 (2005).
8. Bergthorsson, U., Andersson, D.I. & Roth, J.R. Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences USA* **104,** 17004–17009 (2007).
9. Jeffery, C.J. Moonlighting proteins: old proteins learning new tricks. *Trends in Genetics* **19,** 415–417 (2003).
10. Huberts, D.H.E.W. & van der Klei, I.J. Moonlighting proteins: an intriguing mode of multitasking. *Biochimica et Biophysica Acta: Molecular Cell Research* **1803,** 520–525 (2010).
11. Dawkins, R. *Climbing mount improbable.* (Norton, 1996).
12. Renata, H., Wang, Z.J. & Arnold, F.H. Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. *Angewandte Chemie (International Edition)* **54,** 3351–3367 (2015).
13. Shaikh, F.A. & Withers, S.G. Teaching old enzymes new tricks: engineering and evolution of glycosidases and glycosyl transferases for improved glycoside synthesis. *Biochemistry and Cell Biology* **86,** 169–177 (2008).
14. Amitai, G., Gupta, R.D. & Tawfik, D.S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP Journal* **1,** 67 (2007).
15. Hasnaoui-Dijoux, G., Majerić Elenkov, M., Lutje Spelberg, J.H., Hauer, B. & Janssen, D.B. Catalytic promiscuity of halohydrin dehalogenase and its application in enantioselective epoxide ring opening. *ChemBiochem* **9,** 1048–1051 (2008).
16. Cheriyan, M. *et al.* Directed evolution of a pyruvate aldolase to recognize a long chain acyl substrate. *Bioorganic & Medicinal Chemistry* **19,** 6447–6453 (2011).
17. Toscano, M.D., Woycechowsky, K.J. & Hilvert, D. Minimalist active-site redesign: teaching old enzymes new tricks. *Angewandte Chemie (International Edition)* **46,** 3212–3236 (2007).
18. Hochberg, G.K.A. *et al.* A hydrophobic ratchet entrenches molecular complexes. *Nature* **588,** 503–508 (2020).
19. Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F. & Gray, M.W. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63,** 528–537 (2011).
20. Zou, Z. & Zhang, J. Morphological and molecular convergences in mammalian phylogenetics. *Nature Communications* **7,** 12758 (2016).
21. Schulz, L. *et al.* Evolution of increased complexity and specificity at the dawn of form I Rubiscos. *Science* **378,** 155–160 (2022).
22. Pillai, A.S. *et al.* Origin of complexity in haemoglobin evolution. *Nature* **581,** 480–485 (2020).

23. Finnigan, G.C., Hanson-Smith, V., Stevens, T.H. & Thornton, J.W. Evolution of increased complexity in a molecular machine. *Nature* **481,** 360–364 (2012).

24. Goldschmidt, R.B. *The material basis of evolution* (Yale University Press, 1982).

25. Warinner, C., Korzow Richter, K. & Collins, M.J. Paleoproteomics. *Chemical Reviews* **122,** 13401–13446 (2022).

26. Novak, B.J. De-Extinction. *Genes* **9**, 548 (2018).

27. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nature Reviews. Genetics* **13,** 303–314 (2012).

28. Hochberg, G.K.A. & Thornton, J.W. Reconstructing ancient proteins to understand the causes of structure and function. *Annual Review of Biophysics* **46,** 247–269 (2017).

29. Sleator, R.D. Phylogenetics. *Archives of Microbiology* **193,** 235–239 (2011).

30. Darwin, C. & Barrett, P.H. (eds.). *Charles Darwin's notebooks, 1836 - 1844. Geology, transmutation of species, metaphys. enquiries* (Cornell University Press, 1987).

31. Felsenstein, J. *Inferring phylogenies* (Sinauer Association, 2004).

32. Fitch, W.M. Toward defining the course of evolution. Minimum change for a specific tree topology. *Systematic Zoology* **20,** 406 (1971).

33. Bretthorst, G.L. & Jaynes, E.T. (eds.). *Probability theory. The logic of science.* (Cambridge University Press, 2013).

34. Burgie, E.S., Walker, J.M., Phillips, G.N. & Vierstra, R.D. A photo-labile thioether linkage to phycoviolobilin provides the foundation for the blue/green photocycles in DXCF-cyanobacteriochromes. *Structure* **21,** 88–97 (2013).

35. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33,** 3390–3400 (2005).

36. Mascotti, M.L. Resurrecting enzymes by ancestral sequence reconstruction. *Methods in Molecular Biology* **2397,** 111–136 (2022).

37. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).

38. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004).

39. Felsenstein, J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22,** 240 (1973).

40. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25,** 1307–1320 (2008).

41. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4,** 406–425 (1987).

42. Sullivan, J. & Joyce, P. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **36,** 445–466 (2005).

43. Hordijk, W. & Gascuel, O. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* **21,** 4338–4347 (2005).

44. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39,** 783–791 (1985).

45. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556,** 452–456 (2018).

46. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology* **55,** 539–552 (2006).

47. Wheeler, W.C. Nucleic acid sequence phylogeny and random outgroups. *Cladistics* **6,** 363–367 (1990).

48. Eick, G.N., Bridgham, J.T., Anderson, D.P., Harms, M.J. & Thornton, J.W. Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Molecular Biology and Evolution* **34,** 247–261 (2017).

49. Cardona, T., Murray, J.W. & Rutherford, A.W. Origin and evolution of water oxidation before the last common ancestor of the cyanobacteria. *Molecular Biology and Evolution* **32,** 1310–1328 (2015).

50. Möglich, A., Yang, X., Ayers, R.A. & Moffat, K. Structure and function of plant photoreceptors. *Annual Review of Plant Biology* **61,** 21–47 (2010).

51. Anders, K. & Essen, L.O. The family of phytochrome-like photoreceptors: diverse, complex and multi-colored, but very useful. *Current Opinion in Structural Biology* **35,** 7–16 (2015).

52. Rockwell, N.C. & Lagarias, J.C. A brief history of phytochromes. *ChemPhysChem* **11,** 1172–1180 (2010).

53. Ikeuchi, M. & Ishizuka, T. Cyanobacteriochromes: a new superfamily of tetrapyrrole-binding photoreceptors in cyanobacteria. *Photochemical & Photobiological Sciences* **7,** 1159–1167 (2008).

54. Wilson, A. *et al.* A soluble carotenoid protein involved in phycobilisome-related energy dissipation in cyanobacteria. *The Plant Cell* **18,** 992–1007 (2006).

55. Wilson, A. *et al.* A photoactive carotenoid protein acting as light intensity sensor. *Proceedings of the National Academy USA* **105,** 12075–12080 (2008).

56. Gwizdala, M., Wilson, A. & Kirilovsky, D. In vitro reconstitution of the cyanobacterial photoprotective mechanism mediated by the orange carotenoid protein in *Synechocystis* PCC 6803. *The Plant Cell* **23,** 2631–2643 (2011).

57. Bao, H. *et al.* Additional families of orange carotenoid proteins in the photoprotective system of cyanobacteria. *Nature Plants* **3,** 17089 (2017).

58. Boulay, C., Wilson, A., D'Haene, S. & Kirilovsky, D. Identification of a protein required for recovery of full antenna capacity in OCP-related photoprotective mechanism in cyanobacteria. *Proceedings of the National Academy of Sciences USA* **107,** 11620–11625 (2010).

59. Muzzopappa, F., Wilson, A. & Kirilovsky, D. Interdomain interactions reveal the molecular evolution of the orange carotenoid protein. *Nature Plants* **5,** 1076–1086 (2019).

60. Thurotte, A. *et al.* The cyanobacterial fluorescence recovery protein has two distinct activities: orange carotenoid protein amino acids involved in FRP interaction. *Biochimica et Biophysica Acta: Bioenergetics* **1858,** 308–317 (2017).

Chapter II

Original Research Publications

Evidence for an early green/red photocycle that precedes the diversification of GAF domain photoreceptor cyanobacteriochromes

This work is published in:

Priyadarshini, N.[1] & **Steube, N.[1]**, Wiens, D., Narikawa, R., Wilde, A., Hochberg, G.K.A, Enomoto, G. Evidence for an early green/red photocycle that precedes the diversification of GAF domain photoreceptor cyanobacteriochromes. *Photochemical and Photobiological Sciences* **22**, 1415–1427 (2023).

[1] Nibedita Priyadarshini & Niklas Steube have contributed equally to this work and are listed in alphabetical order.

The online version of the publication is open access and contains all supplementary material, available at https://doi.org/10.1007/s43630-023-00387-4.

Own contribution:

Niklas Steube gathered sequence data, performed all phylogenetic analyses and ancestral protein sequence reconstructions. He further performed AlphaFold2 structure prediction analyses of the reconstructed proteins, designed figures, and contributed in discussing the data and drafting the manuscript.

Co-author's contribution:

GKAH and GE designed research; NP, NS, DW, and GE performed research; all authors analyzed data; all authors contributed to writing the paper.

Abstract

Phytochromes are linear tetrapyrrole-binding photoreceptors in eukaryotes and bacteria, primarily responding to red and far-red light signals reversibly. Among the GAF domain-based phytochrome superfamily, cyanobacteria-specific cyanobacteriochromes show various optical properties covering the entire visible region. It is unknown what physiological demands drove the evolution of cyanobacteriochromes in cyanobacteria. Here, we utilize ancestral sequence reconstruction and biochemical verification to show that the resurrected ancestral cyanobacteriochrome proteins reversibly respond to green- and red-light signals. pH titration analyses indicate that the deprotonation of the bound phycocyanobilin chromophore is crucial to perceive green light. The ancestral cyanobacteriochromes show only modest thermal reversion to the green light-absorbing form, suggesting that they evolved to sense the incident green/red light ratio. Many cyanobacteria can utilize green light for photosynthesis using phycobilisome light-harvesting complexes. The green/red sensing cyanobacteriochromes may have allowed better acclimation to changing light environments by rearranging the absorption capacity of the phycobilisome through chromatic acclimation.

Introduction

Most light-dependent cellular responses are controlled by photoreceptors which sense light and then trigger down-stream signal transduction events[1]. Members of the phytochrome superfamily of photoreceptors covalently bind a linear tetrapyrrole (bilin) molecule as a chromophore to a cysteine (Cys) residue of the protein[2,3]. The configuration of the bound bilin chromophore reversibly interconverts between *15Z* and *15E*, corresponding to the two isomers at the C15=C16 double bond[4] (Fig. S1). These two states of the chromophore often result in different optical properties, enabling the proteins to sense two different colors of light, in most cases red and far-red. The reversible photochromicity allows the photoreceptor to perceive the ratio of two wavelengths of the incident light. Many phytochromes show thermal reversion (dark reversion), reverting from *15E* to *15Z* without light absorption. Thermal reversion is a temperature-dependent process, and therefore the same photoreceptor integrates light and temperature signals[5,6]. A fast dark reversion of a photoreceptor indicates that the protein senses the intensity of the incident light rather than the ratio of the two wavelengths[7-10].

Within the phytochrome superfamily, cyanobacteriochromes (CBCRs) are a distinct class of minimal photoreceptors[11,12], which only need a single GAF (cGMP phosphodiesterase, adenylyl cyclase, and FhlA) domain to sense light genuinely. This contrasts with other phytochrome members that strictly require at least another neighboring PHY domain for genuine light perception[2,3]. The functional light sensing module of canonical phytochromes features a typical PAS-GAF-PHY tri-domain architecture, with the exception of some members lacking the PAS domain (knotless phytochromes) that are closely related to CBCRs[2,3]. Phytochromes are widespread among eukaryotes and bacteria, whereas CBCRs are found exclusively in cyanobacteria, a group of photoautotrophic bacteria performing oxygenic photosynthesis. Through a process of gene duplication and domain shuffling, CBCRs have evolved a remarkable diversity in their absorption characteristics and thermal reversion kinetics[7,13-16], making them a promising scaffold to develop a new generation of optogenetic tools[10,17,18]. Depending on their properties, CBCRs control a diverse range of physiological processes in cyanobacteria[19]. Green/red sensing CBCRs with slow reversion kinetics, including the first discovered CBCR *RcaE*, are used to adjust the relative amounts of red and green absorbing photosynthetic pigments (phycocyanin and phycoerythrin, respectively) in phycobilisomes during chromatic acclimation by sensing the ratio of green and red wavelengths[15,20-22]. Blue/green sensing CBCRs, on the other hand, are considered to be used to detect shading by other cells in cyanobacterial mats[23,24].

However, the original function of CBCRs remains unknown. We have previously speculated that blue/green perceiving CBCR-mediated cell shade sensing might be the ancestral function of these photoreceptors[23] because blue/green photochemistry is unique to CBCRs and should be more efficient than red/far-red phytochromes in an upper region of a microbial mat, where blue light diminishes while green, red, and far-red light are still available[25]. Further, early-branching cyanobacteria such as *Gloeobacter violaceus* PCC 7421 and *Anthocerotibacter panamensis*[26] only possess potential relatives of this kind of blue/green perceiving CBCRs based on sequence similarity, although they have not yet been characterized biochemically. However, the phylogenetic history of CBCRs is very complex, including frequent gene and domain duplications, making this question hard to resolve. It is difficult to make unambiguous predictions about the properties of the last common ancestor (LCA) of all CBCR GAF domains using existing phylogenies, because not enough GAF domains have been characterized biochemically, and their relative branching order remains uncertain[27].

Here, we used ancestral sequence reconstruction[28] to experimentally determine the photochemistry of the LCA of all extant CBCRs. We show that ancient CBCR proteins most likely sensed the ratio of green/red incident light, but not blue/green light. This inference is robust to alternative hypotheses about the exact branching order within CBCR GAF domains that is hard to resolve. Our results suggest that the first CBCR was likely used by cyanobacteria to tune the relative abundances of red and green light-absorbing pigments in response to changes in the incident light. The stunning diversity of colors sensed by extant CBCRs nowadays, therefore, may have evolved from an ancient CBCR most likely used for chromatic acclimation.

Results

Ancestral sequence reconstruction of cyanobacteriochromes (CBCRs)

To investigate the characteristics of the earliest CBCRs, we first used maximum likelihood (ML) phylogenetics and ancestral sequence reconstruction to infer the most likely GAF domain sequence of the LCA of all extant CBCRs. To do this, we used HMMER to identify all CBCR GAF domains in 30 cyanobacterial species that span the entire known species diversity. We inferred a maximum likelihood phylogeny of 575 CBCR GAF domains. Although it is not yet clear which family of phytochromes evolved first, it is uncontroversial that knotless phytochromes form a closely related sister group to CBCRs[29]. Thus, we used 45 cyanobacterial knotless phytochrome GAF domains as the outgroup to root our tree.

The tree clearly separates the GAF domains of all cyanobacterial knotless phytochromes from the ones of all CBCRs on our tree (Fig. 1a, Fig. S2). Beyond that, the phylogeny of the CBCR domains was extremely difficult to resolve. Our maximum likelihood tree did not contain any well-supported monophyletic groups of CBCR domains that clearly originated from gene duplication or domain-swapping events. CBCR domains are grouped loosely by domain architecture of the full-length proteins they are found in, but even these architectures vary substantially among GAF domains that group closely together. Mapping known CBCR color-sensing characteristics on the tree did not reveal an obvious pattern or a clear inference for the ancestral color. The earliest branching CBCRs on the tree presented here are green/red, red/orange, and green/blue receptors. The clade containing green/blue receptors connects to the root via a long branch, so its placement may result from a long branch attraction artifact (Fig. 1a, Fig. S2).

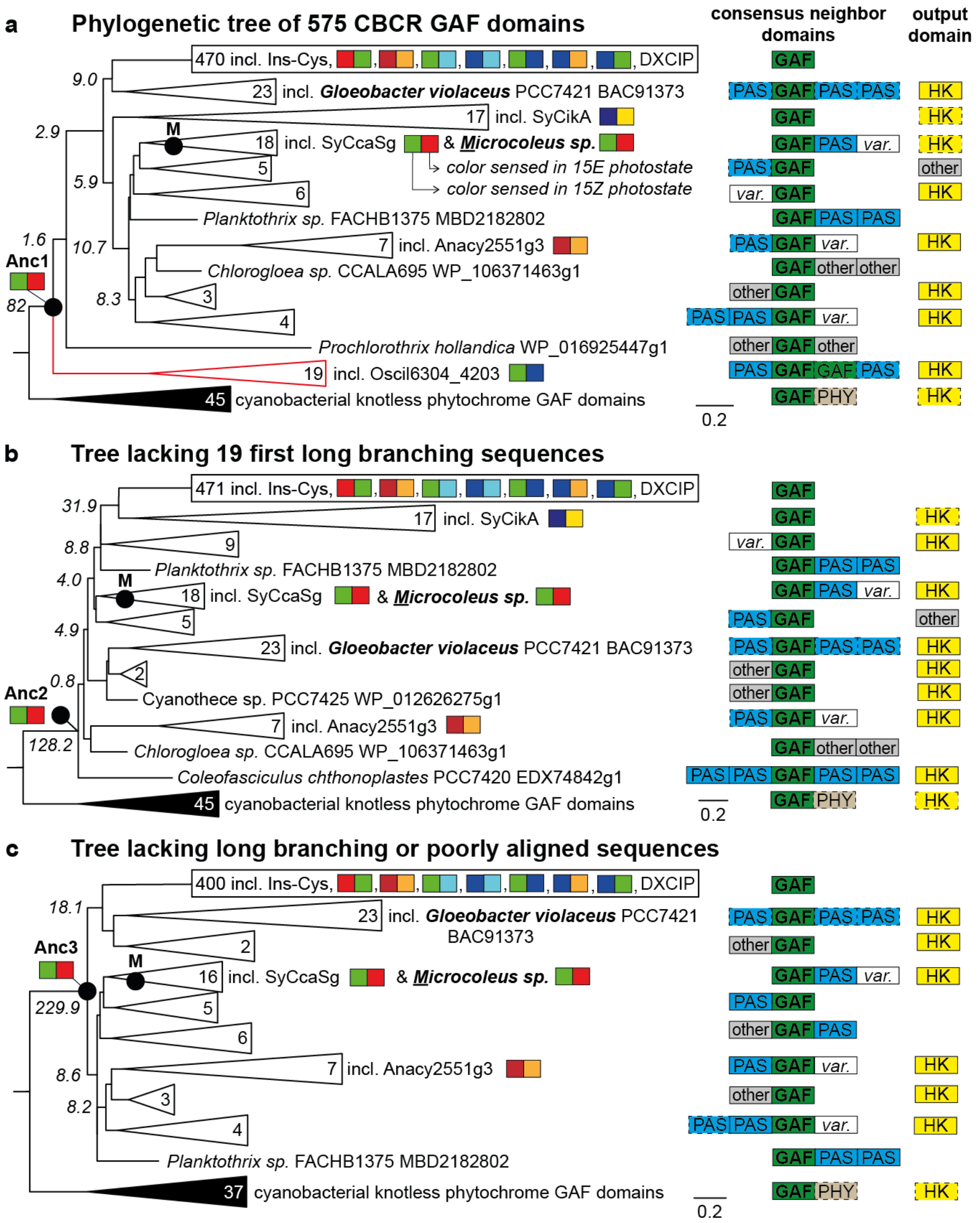
Our phylogenetic tree implies that the exact branching order of CBCR GAF domains is not resolvable with current methods, making inferences about the LCA impossible by comparing only the absorption/emission spectra of extant CBCRs. We reasoned that we might still gain some insights into its potential properties by ancestral sequence reconstruction, even if the topology of our ML tree could be wrong within the CBCR domains. Ancestral sequence reconstruction infers the likely sequence at internal nodes of the tree, given the tree topology, alignment, and a model of sequence evolution[28]. We reasoned that we could use this technique to test whether our different trees imply any consistent emission/absorption properties that are robust to phylogenetic uncertainty. All basal internal branches on our tree are short and poorly supported. Under such circumstances, it is possible that such errors do not affect reconstructions at functionally important sites (for which the signal should be strong)[30].

To test if there is any phylogenetic signal for a particular color sensing of the LCA on our tree phylogeny, we decided to use ancestral protein resurrection to test biochemically which color our tree implies. To do this, we inferred the sequence of the LCA of all CBCRs on our tree, resurrected the ancient GAF domain, and characterized it biochemically, as reported below. To determine if those characteristics strongly depend on the exact branching order of the tree or if they are robust to slight rearrangements of the poorly resolved branching order within CBCR GAF domains, we decided to infer two additional trees. For one, we only removed the first clade of long branching green/blue receptors and re-inferred the tree (Fig. 1b). For a third tree, we additionally removed sequences that were only poorly aligned or very long branching on our first tree (Fig. 1c). The two additional trees did not improve on the unresolved branching pattern inside the CBCRs, but had slight rearrangements near the root. Notably, in all three trees, the single GAF domain found in *Gloeobacter violaceus* PCC 7421 (the earliest branching cyanobacterial species on our trees) branched near the root. Furthermore, far-red/orange *Ancy2551g3* and green/red *SyCcaSg* always appeared as early branching among the known characterized CBCR GAF domains. All three trees would be incorrect in the exact branching order within CBCR GAF domains. We, therefore, view the ancestral sequences we inferred from them not as a historically accurate inference, but simply as a test for whether there is any residual phylogenetic signal for the color of the LCA of all CBCRs that may be robust to slight rearrangements of branches near the root.

We inferred the most likely amino acid sequences of the LCA of extant CBCR GAF domains on all three topologies (Anc1–Anc3) to an average posterior probability of 0.81, 0.92, and 0.94, respectively (Fig. S3). The ancestral sequences all contained the conserved “first” cysteine that binds the bilin chromophore in extant CBCRs as expected but differed at between 37 and 44 out of 142 total residues (Fig. 1d, Fig. S4). To further validate our findings, we attempted to characterize CBCR GAF domains of early branching extant species that only have short evolutionary distance (in branch lengths on our trees) to the reconstructed ancestral CBCR GAF domain sequences and review if their biochemical properties match the suggested ones of the ancestors.

CBCR GAF domains are located on a variety of proteins ranging from single domain up to multi-domain proteins that often contain several GAF domains. Some

GAF domains function on their own as a CBCR, and others belong to other phytochromes that are strictly dependent on adjacent domains for genuine light perception. Large evolutionary distances between GAF domains on the same protein indicate early domain duplications or frequent horizontal transfer events between cyanobacterial species (Fig. S5). To estimate the most probable domain architecture of the ancestral CBCR protein, we further compared the neighbor and output domains of the corresponding full-length proteins of CBCR GAF domains on all our trees. We found PAS domains mandatory in distantly related canonical phytochromes as the most abundant neighbors, and histidine kinase/HATPase domains as the most prominent output domains in early branching CBCRs on our trees (Fig. 1). The trees presented here, thus, indicate that the LCA of all CBCRs was probably encoded on a phytochrome-like multidomain protein and transduced its signal to a histidine kinase domain.


a
Phylogenetic tree of 575 CBCR GAF domains
consensus neighbor domains
output domain
470 incl. Ins-Cys, DXCIP
23 incl. Gloeobacter violaceus PCC7421 BAC91373
17 incl. SyCikA
M
18 incl. SyCcaSg & Microcoleus sp.
5
6
color sensed in 15E photostate
color sensed in 15Z photostate
Planktothrix sp. FACHB1375 MBD2182802
7 incl. Anacy2551g3
Chlorogloea sp. CCALA695 WP_106371463g1
3
4
Prochlorothrix hollandica WP_016925447g1
19 incl. Oscil6304_4203
45 cyanobacterial knotless phytochrome GAF domains
9.0
2.9
5.9
1.6
10.7
Anc1
82
8.3
0.2
GAF
PAS
var.
other
PHY
HK
b
Tree lacking 19 first long branching sequences
471 incl. Ins-Cys, DXCIP
17 incl. SyCikA
9
Planktothrix sp. FACHB1375 MBD2182802
18 incl. SyCcaSg & Microcoleus sp.
5
23 incl. Gloeobacter violaceus PCC7421 BAC91373
2
Cyanothece sp. PCC7425 WP_012626275g1
7 incl. Anacy2551g3
Chlorogloea sp. CCALA695 WP_106371463g1
Coleofasciculus chthonoplastes PCC7420 EDX74842g1
45 cyanobacterial knotless phytochrome GAF domains
31.9
8.8
4.0
4.9
0.8
Anc2
128.2
0.2
c
Tree lacking long branching or poorly aligned sequences
400 incl. Ins-Cys, DXCIP
23 incl. Gloeobacter violaceus PCC7421 BAC91373
2
16 incl. SyCcaSg & Microcoleus sp.
5
6
7 incl. Anacy2551g3
3
4
Planktothrix sp. FACHB1375 MBD2182802
37 cyanobacterial knotless phytochrome GAF domains
18.1
Anc3
229.9
8.6
8.2
0.2

d **Sequences of examined reconstructed ancestral and extant GAF domains**

```
Anc1  DLEEILNTTVTEVRQFLQTD RVLIYRFQPDGSGTVIAESV NPGWPSLLGQTFPADCIPPE YLEQYRQGRLRSISD 75
Anc2  DLEEILNTTVTEVRQLLQCD RVLIYRLWPDGTGSVVAEAV VPGWPAILGQTFPEEVFPPE CHQLYCQGRIRAIAD 75
Anc3  NLEEILNTTVTEVRQFLQCD RVLIYRFWPDGSGSVVAEAV APGWPSILGQTFPEEVFPEE YHELYCQGRIRAIAD 75
M     NLEEILNTTVTEVRQFLQTD RVLIYRLWPNGTGSAVTEAV VPGWPTVLGRTFPEEVFPLE SHKAYCQGRILAISD 75
                ▼1st Cys                     Hallmark Asp position ▲ ▲2nd Cys position

Anc1  IEAAN LSPCHVELLQQMQVKSNLVV PILQQDQLWGLLVAHHCQSP RQWSPQELQLLQQVANQIAI AI 142
Anc2  VEQDN ISPCLVEFLQQFGVKSKLVV PILQKEKLWGLLIAHHCSSP RQWQPFEIELLQQLATQLAI AI 142
Anc3  VEQAN ISPCHVEFLQQFGVKANLVV PILQKDQLWGLLIAHQCSGP RQWQPFEIELLQQLATQIAI AI 142
M     VEQSI VLPCLVEFVQQFGVKAKLVV PILQDDTLWGLLIAHHCSSP RQWQPLEIDLLQSLATQLAI AL 142
```

Fig. 1| Ancestral CBCR GAF domain reconstruction on ML phylogenies. a-c, Maximum Likelihood phylogenetic trees of cyanobacterial GAF domains used for ancestral sequence reconstruction. Numbers labeling clades denote the quantity of taxa. Colored squares highlight biochemically characterized domains and the colors they sense. “Ins-Cys” and “DXCIP” denote families sensing various colors. “M” indicates the extant, early-branching CBCR GAF domain of *Microcoleus* sp. FACHB1 MBD2125673 that we characterized. The clade of 19 first branching sequences shown in red was deleted for tree B. Node support is shown as approximate likelihood test statistics in italics. Scale bar: 0.2 average substitutions per site. Consensus neighbor and output domains of corresponding full-length proteins are shown to the right of the trees with domains that only appear in most of the proteins with dashed outlines. *var.*, variable domains. other, conserved domains other than PAS (Per-Arnt-Sim), PHY (phytochrome-specific domain) or HK (histidine kinase/HATPase). **d,** Amino acid sequences of the extant (M) and reconstructed ancestral GAF domains (Anc1-3). Arrows point positions important for color sensing in extant CBCRs, and states are red if conserved and blue if not.

Signal for a green/red photocycle in all ancestral CBCR GAF domains

We next determined the photochemical properties of the ancestral CBCR GAF domains. We expressed and purified the three ancestral sequences as recombinant N-terminal His-tagged proteins from *E. coli* harboring a biosynthesis plasmid for the chromophore phycocyanobilin (PCB). The Zn^{2+}-enhanced fluorescence of the purified proteins in an SDS-PAGE gel confirmed the covalent attachment of a bilin chromophore to the apoproteins (Fig. S6)[31]. The absorbance spectra of the purified holo-proteins showed spectral changes upon illumination with blue (λ_{max} = 448 nm), green (λ_{max} = 514 nm), and red light (λ_{max} = 635 nm). Irradiation with UV (λ_{max} = 355 nm) and far-red light (λ_{max} = 731 nm) did not affect the spectra. All ancestral proteins exhibited reversible photoconversion between green (Pg) and red (Pr) absorbing forms (Fig. 2). The bound chromophore species and its configuration were determined using acid denaturation spectra. The acid-denatured red-irradiated state (*i.e.*, Pg) showed a peak at 662 nm and the green-irradiated state (*i.e.*, Pr) at 585 nm, in agreement with *15Z* and *15E* forms of the covalently bound PCB, respectively (Fig. S7)[32], indicating that Pg carries *15Z* PCB whereas Pr has *15E* PCB. The ^{15Z}Pg state showed absorption maxima between 515 nm and 540 nm, and the ^{15E}Pr state between 600 nm and 656 nm for all the ancestral proteins (Fig. 2, Tab. 1). For Anc2 and Anc3, irradiation with red (λ_{max} = 635 nm) resulted in almost complete conversion to the ^{15Z}Pg form. For Anc1, we did not yield an apparently homogeneous population of ^{15Z}Pg by red light irradiation, probably due to the significant overlap of the absorption spectra of the two photo-states (Fig. 2, Fig. S7). The additional incubation of Anc1 overnight in the dark at room temperature allowed a seemingly complete conversion to ^{15Z}Pg (Fig. S7). Irradiation with blue (λ_{max} = 448 nm) and green (λ_{max} = 514 nm) rendered almost complete conversion to the ^{15E}Pr state for Anc1 and Anc2. For Anc3, green irradiation resulted in partial conversion. The almost complete conversion was achieved upon blue irradiation, probably due to its good separation from the counteracting red region (Fig. 2, Fig. S7). Although blue light could induce photoconversion, we characterize the ancestral proteins as green-light sensors because the peak wavelengths of the absorption spectra and the difference spectra both fall into the green-light region (Fig. 2D).

We attempted to characterize further CBCR GAF domains of early branching extant species with a short evolutionary distance to the reconstructed ancestors on our trees, namely *Chlorogloea* sp. CCALA 695 WP_106371463.1, *Oscillatoria* sp.

PCC 10802 WP_082218260.1, and *Microcoleus* sp. FACHB1 MBD2125673 WP_190776511.1. As we were not able to heterologously express sufficient amounts of the first two, we characterized the CBCR GAF domain of *Microcoleus* sp. with an evolutionary distance between 0.31 and 0.67 on our trees, and found the same green/red perception as in the ancestral domains (Fig. S8).

Taken together, although the spectral shapes are distinct among the three ancestral and the extant CBCR GAF domains, our results show a phylogenetic signal for a green/red photocycle in the LCA of all CBCRs, regardless of the exact branching order of basal CBCRs.

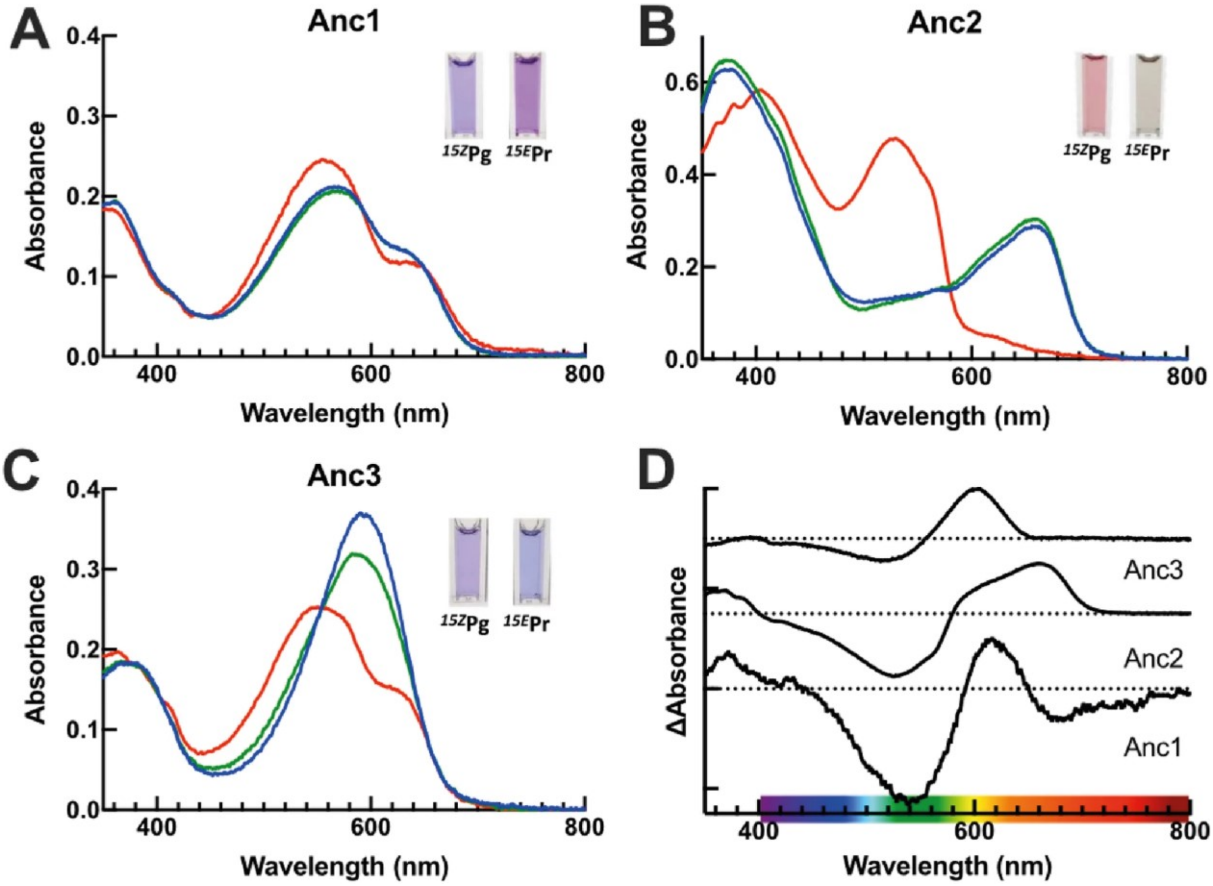


Fig. 2| Absorption and difference spectra of the purified ancestral proteins. A-C, Absorption spectra of the ^{15Z}Pg (red line), and of the ^{15E}Pr form (blue and green lines) of Anc1-3. The ^{15Z}Pg form was achieved by irradiation with red, the ^{15E}Pr form by either irradiation with blue or green for one minute. **D,** Normalized photochemical difference spectra obtained by subtracting the absorption spectra of the ^{15Z}Pg from those of the ^{15E}Pr form of Anc1-3. Difference spectra were normalized to the red photoproduct peak, and are vertically shifted for clarity. **A-C insets,** The difference in the color of the ^{15Z}Pg and the ^{15E}Pr forms of Anc1-3 in solution at pH 7.5. All experiments were performed at room temperature.

Tab. 1| Wavelengths of the absorbance peak maxima and the half-lives of thermal reversion of ancestral CBCR proteins at room temperature.

Protein	$\lambda_{\text{max}, 15Z}$ (nm)	$\lambda_{\text{max}, 15E}$ (nm)	$\lambda_{\text{max}, 15Z, \text{denatured}}$ (nm)	$\lambda_{\text{max}, 15E, \text{denatured}}$ (nm)	half-life (min)	reference
Anc1 WT	540	610	663	589	233	this work
Anc2 WT	525	656	663	583	180	this work
Anc3 WT	515	600	667	589	310	this work
Anc1 C56V	541	621	n.d.	n.d.	n.d.	this work
Anc1 A54D	535	620	n.d.	n.d.	n.d.	this work
Anc2 E54D	525	660	n.d.	n.d.	n.d.	this work
Anc3 E54D	515	602	n.d.	n.d.	n.d.	this work
Cph1-PCB	n.d.	n.d.	669	573	n.d.	32
TePixJ-PVB	n.d.	n.d.	600	507	n.d.	32

The peak wavelengths were calculated using the difference spectra upon reversible photoconversion. n.d., not determined.

PCB was the chromophore in ancestral CBCRs

Although most CBCRs incorporate PCB, some CBCRs can bind biliverdin IXa (BV) as the chromophore with variable specificity[33,34]. To determine the efficiency of BV incorporation by the ancestral proteins, we expressed all of them with a BV biosynthesis plasmid in *E. coli* and purified them (Fig. S6). Acid denaturation spectra confirmed the attached chromophore to be BV with the denatured ^{15Z}Pg peaking at around 700 nm (Fig. S9)[34]. All ancestral proteins showed slight photoconversion with BV as the chromophore upon irradiation with both green and red light. However, for Anc1 and Anc2, neither lights were sufficient to cause complete photoconversion to either *15E* or *15Z* photo-states (Fig. S9). Red irradiation caused a complete conversion of Anc3-BV to the *15Z* photo-state. However, a complete conversion to the *15E* photo-state was not achieved by green irradiation. These data suggest that the ancestral CBCRs may have been able to bind to both, PCB and BV, but that photoconversion may have been efficient with PCB. Specificity for BV would then be a derived trait of some crown-group CBCRs[33]. This is consistent with cyanobacterial knotless phytochromes in the outgroup, also being specific for PCB[35,36]. Besides, PCB is one of the prosthetic groups of the phycobiliproteins of the photosynthetic antenna complex and is much more abundant than BV in cyanobacterial cells[37].

CBCR GAF reconstructions suggest a function as a sensor of the spectral ratio via a protochromic photocycle

We next asked whether the heterologously expressed ancestral proteins sensed the intensity of green or red light rather than the red/green ratio. To determine this, we measured their rates of thermal reversion. Fast thermal reversion leads to short-lived photoproducts regardless of any counteracting light. Therefore, the population of the photoproduct only depends on the intensity of light that excites the dark state[7-10]. In contrast, slow thermal reversion allows the formation of long-lived photo-states and therefore supports sensing of the ratio of two different wavelengths. All three ancestral proteins underwent slow thermal reversion from ^{15E}Pr to ^{15Z}Pg in the dark at room temperature (Fig. S10): The half-lives for the thermal reversion in the dark at room temperature ranged between 180 min and 310 min (Tab. 1), comparable to the related knotless phytochromes[35]. These half-lives are much longer than those of known intensity sensing CBCRs, which revert within the range of several seconds[7-10]. Our results, therefore, indicate that the LCA of all CBCRs likely sensed the ratio of green to red incident light rather than the intensity of these wavelengths.

Extant green/red light-sensing CBCRs adopt a protochromic photocycle[15,38]. The conjugated π-system of the bilin chromophore of the green/red CBCRs is deprotonated with a lower pKa value in the *15Z* state to absorb green light, whereas it is protonated with a higher pKa value in the *15E* state to absorb red light. To assess whether this was also the ancestral photocycle mechanism in CBCR GAF domains, we performed pH titration analysis for the three ancestral proteins.

Anc1–3 showed a decrease in absorption in the red-light region (600–660 nm) and an increase in green-light absorption (520–540 nm) at higher pH conditions (Fig. 3, Fig. S11). At lower pH conditions, red-light absorption increased and green-light absorption decreased, except for Anc2 *15Z*, which showed stable green-light absorption under the tested pH conditions. The absorption changes were fitted with one titrating group of the Henderson–Hasselbalch equation to estimate pKa[15]. The pKa values of the *15Z* chromophore are lower than those of *15E*, indicating that the *15Z* chromophore has a lower affinity to protons (Tab. 2). The difference in pKa values between *15Z* and *15E* was the smallest in Anc1 (Tab. 2), which may be consistent with its poor spectral shift upon photoconversion under the standard pH condition of 7.5 (Fig. 2). The much lower pKa of Anc2 *15Z* may be due to the leucine residue next to the chromophore-binding cysteine, which is important for stabilization of the deprotonation of the chromophore[15,39]. These results suggest that a photochromic photocycle similar to that of extant green/red CBCRs may have been the ancestral photo-switching mechanism.

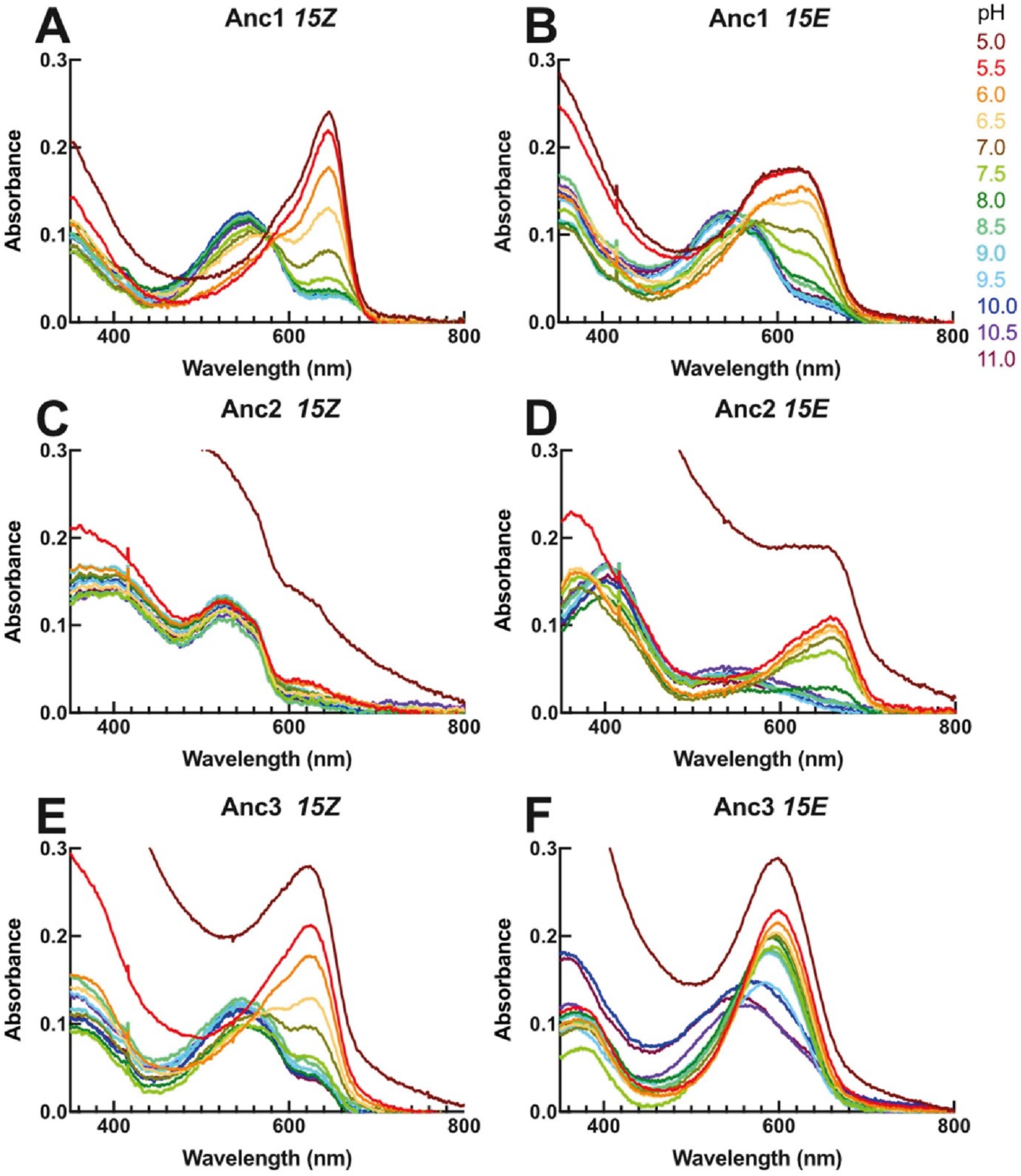


Fig. 3| Protochromic absorption spectra changes of the ancestral proteins. A-F, pH-dependent absorbance spectra of Anc1-3 with the configuration of *15Z* (**A, C, E**) or *15E* (**B, D, F**), measured in buffers with pH between 5.0 (dark red) and 11.0 (dark purple) in 0.5 pH steps. Increased scattering was observed at lower pH of 5.0 and 5.5, probably due to partial protein aggregation. For the analysis, samples were irradiated to obtain homogenous *15Z* and *15E* photo-states, followed by mixing with 1 M buffers of different pH in 1:4 ratio and immediate measurement of absorption spectra. Note that the homogenous *15Z* of Anc1 was prepared by overnight incubation of the protein in the dark.

Tab. 2| The estimated pKa values of the ancestral CBCR proteins.

Protein/configuration	pKa	absorption peaks for fitting (nm)	R^2
Anc1 WT/*15Z*	6.54	650	0.9994
Anc1 WT/*15E*	7.22	635	0.9936
Anc2 WT/*15Z*	< 5.0	n.a.	n.a.
Anc2 WT/*15E*	7.57	670	0.9613
Anc3 WT/*15Z*	6.59	630	0.9859
Anc3 WT/*15E*	9.35	610	0.9501
Anc1 C56V/*15Z*	6.77	650	0.9959
Anc1 C56V/*15E*	7.46	635	0.9887
Anc1 A54D/*15Z*	6.57	650	0.9941
Anc1 A54D/*15E*	7.61	635	0.9913
Anc2 E54D/*15Z*	< 5.0	n.a.	n.a.
Anc2 E54D/*15E*	7.75	670	0.9955
Anc3 E54D/*15Z*	6.58	630	0.9653
Anc3 E54D/*15E*	9.48	610	0.9697

The pKa values were calculated using the data of the absorption changes in the pH titration experiments in Fig. 3; n.a., not applicable.

The amino acids aligned at the conserved CBCR hallmark residues do not control the green/red photocycle

Lastly, we sought to gain insights into the molecular mechanisms of color tuning of the reconstructed CBCR proteins relative to canonical red/far-red phytochromes. We first focused on what allows deprotonation of the chromophore. In canonical phytochromes, the chromophore is protonated in both photo-states[35,38,40,41]. The protonated state is stabilized by a conserved aspartate (Asp) residue at position 54 (numbering of the amino acid is based on the multiple sequence alignment (Fig. S4)) that forms a hydrogen bond network with the nitrogen atoms of the B and C pyrrole rings of the chromophore[42-45]. The resurrected CBCR ancestral proteins feature either alanine or glutamate residue at this position, suggesting that the substitution of Asp to a different amino acid might have allowed the deprotonation of the chromophore. To test this hypothesis, we mutated this site to Asp in all three ancestral proteins, mimicking the situation in canonical phytochromes and most CBCRs. We then determined whether the deprotonation of the chromophore was affected. Surprisingly, green-light absorption and deprotonation were unaffected in all three mutants (Tab. 1+2, Fig. S12+13). This suggests that the loss of the protonation-stabilizing Asp was neither essential for the evolution of a deprotonated chromophore in the *15Z* photo-state nor for green-light absorption.

Finally, we investigated the influence of another site – the so-called ‘second cysteine’ at position 56 that is known to influence spectral tuning in extant CBCRs. CBCRs containing this Cys residue form a thioether linkage with the C10 position of the bilin chromophore[46]. The disruption of the π-conjugated system at the C10 position leads to absorption in the UV-to-blue region[14,47]. The covalent bond formation between the chromophore and the second Cys can be reversibly induced by the light-induced conformational change of the chromophore and the protein. Some 2nd-Cys-containing proteins retain the covalent bond in both *15Z* and *15E* states. The evolution of this second Cys could have contributed to the spectral properties that distinguish CBCRs from canonical phytochromes. However, the predicted ancestral sequences disagree with the presence of the second Cys in the LCA of all CBCRs: only Anc1 harbors the second Cys residue, whereas Anc2 and Anc3 have a valine at this position (Fig. 1d, Fig. S4). Although all three proteins have a green/red photocycle, this introduces ambiguity about whether the second Cys played an essential role in the evolution of the green/red photocycle. The function of this cysteine may depend on the specific context of the protein, such as the neighboring amino acid residues, although the second Cys is functional in many proteins from different lineages within CBCRs[47]. To address this issue, we mutated the Cys at position 56 of Anc1 to valine (identical to the state in Anc2 and Anc3) and tested for differences in spectral properties. The mutation only slightly elevated the absorbance in the red region compared to the green one of both *15E* and *15Z* photo-states, but without affecting the absorption maxima (Tab. 1+2, Fig. S12+13). This confirms that a green/red photocycle was likely present in the LCA of all CBCR GAF domains, regardless of the presence of the second cysteine in the ancestral protein.

Discussion

The first CBCRs could have functioned in chromatic acclimation

Our results suggest that the LCA of extant CBCRs may have functioned as a green/red light sensor with slow thermal reversion that used a protochromic photocycle similar to that of extant green/red sensing CBCRs. However, we caution that this inference is based on trees with unresolved and presumably incorrect topologies within the CBCRs. The fact that we observed similar properties on three different topologies is encouraging, suggesting that the signal for a green/red photocycle may persist independently of the exact topology. However, biases in the data systematically could favor incorrect topologies that then lead to ancestors with misleading biochemical properties[48]. A green/red photocycle might be the genetically simplest one, and we may observe it because our reconstructions fail to correctly incorporate all states necessary to produce any other kind of photochemistry. In light of these caveats, we do not exclude that the LCA of all CBCR GAF domains had different characteristics.

It is unlikely that more CBCR GAF sequences would improve our inference in the future. Fundamentally, we are limited by the small size (~ 140 aa) and fast evolution of CBCR GAF domains. The complex architecture of CBCR GAF domain-containing proteins further complicates the phylogeny of these proteins. Our trees must contain gene duplications of entire CBCR GAF domain-containing proteins, internal duplications that produce proteins containing two or more CBCR GAF domains, and possibly horizontal transfers, domain fusions, and gene conversion events between adjacent CBCR GAF domains. This makes the gene trees of these domains extremely difficult to interpret. Solving this problem will likely require inferring the histories of other domains in CBCR GAF domain-containing proteins and using reconciliation approaches to infer a global history of how CBCR GAF domains were added and removed from different proteins.

An ancestral green/red photocycle is, however, also likely in the light of ecological relevance. What might have been the physiological function of green/red sensing ancestral CBCRs? The first discovered CBCR, *RcaE*, is a green/red sensing protein as the regulator of chromatic acclimation[15,22]. One plausible answer upon comparison with such extant CBCRs with similar photocycle suggests their involvement in regulating the relative amounts of red-absorbing phycocyanin and green-absorbing phycoerythrin in phycobilisomes during chromatic acclimation[21]. This implies that the LCA of all extant cyanobacteria, in which the here identified ancestral

GAF domain would have existed, already possessed phycoerythrin. The Gloeobacterales (the earliest diverging clade of cyanobacteria) usually possess phycoerythrin, suggesting that the pigment has an ancient origin[26,37,49] and that the ability for chromatic acclimation already existed in the earliest cyanobacteria. The analysis of neighboring domains further supports this hypothesis as the extant known chromatic acclimation regulators harbor an additional PAS domain and a histidine kinase as the output domain[20]. It is of note that extant green/red CBCRs also regulate other types of chromatic acclimation, such as controlling the relative amounts of the yellow-green-absorbing phycoerythrocyanin protein or a rod-membrane linker *CpcL* protein, which assembles a photosystem I-specific phycobilisome only in green light[20,50]. Thus, green/red light sensing could be crucial even for cyanobacterial strains lacking green-absorbing phycoerythrin.

Chromatic acclimation was likely important to early cyanobacteria, as a current analysis points to them having lived in sessile microbial mats[51]. In these environments the availability of different wavelengths of light can change dramatically and rapidly across minute distances, depending on the depth of the cell in the mat or the composition of the overlying cells[23].

Tuning of the chromophore towards green/red sensing

Based on our current work, we can speculate about the genetic mechanism responsible for the evolution of the CBCR's green/red light sensitivity from red/far-red sensing canonical phytochromes. If the green/red photocycle was ancestral to all CBCRs, two changes must have occurred relative to canonical phytochromes: the shift of the *15Z* state from red to green, and that of the *15E* state from far-red to red-light absorption.

In the resurrected CBCR ancestral proteins, the *15Z* state is deprotonated. This is different from phytochromes, in which the bilin chromophore is protonated in both photo-states[35,38,40,41], implying that deprotonation of the chromophore is important for green-light absorption. The ancestral proteins all lack the conserved Asp, which is allegedly important for the stabilization of the protonated state in phytochromes and CBCRs[44,45], suggesting that this substitution may have allowed for deprotonation. The side chain of the Asp residue is involved in the hydrogen bond network with the bilin chromophore in CBCRs[43,52], whereas it is generally oriented toward the outside of the chromophore-binding pocket in phytochromes[53,54]. The AlphaFold2 prediction of the structures of the Anc proteins suggests that the amino acids at the hallmark Asp position could form the hydrogen bond network with the chromophore (Fig. S14). However, introducing the Asp back into the ancestral photoreceptors does not abolish deprotonation, implying the involvement of other factors for deprotonation of the chromophore.

In addition, observations from extant CBCRs and phytochromes suggest that deprotonation alone is likely insufficient to yield green-light absorption: the cyanobacterial canonical phytochrome *Cph1* exhibits a pKa of ~ 9.0 in the *15Z* and *15E* photo-states to stabilize the protonated chromophore. Increasing the solvent pH induces a decrease in red-light absorption by *Cph1* but does not cause an increase in green-light absorption[55]. The red/green CBCR *AnPixJg2* retains the protonated chromophore even at the green-absorbing state, and artificial deprotonation does not affect the green absorption[56]. This suggests that green absorption requires additional amino acid substitutions affecting the light wavelength absorbed by the deprotonated chromophore.

The *15E* state is also hypsochromically shifted from far-red to red absorption. This could have occurred through the loss of the adjacent PHY domain from an ancestral phytochrome-like precursor. Such truncations led to a blue shift of the far-red absorbing state of extant phytochromes[36,42,57]. Another suggested tuning mechanism is the “second” Cys, which is found near the chromophore and is known to influence the absorption properties of proteins from various lineages of CBCR GAF domains[14,47,58,59]. However, the reconstructed ancestral proteins vary in the amino acid at that position; Anc1 has a Cys, whereas Anc2 and Anc3 both have valine. Although the AlphaFold2 prediction locates the second Cys near the C10 of the chromophore (Fig. S14), mutating this cysteine in Anc1 has essentially no effect on optical properties, suggesting that in the LCA of all CBCR GAF domains this site is likely not involved in color tuning. Further exploration would be necessary to shed light on the exact genetic mechanism that transformed a likely red/far-red sensing phytochrome into a green/red sensing CBCR.

The genetic basis of CBCRs may have diversified from an ancestral green/red light sensor

Our results hint at how the remarkable diversity of colors found in extant CBCRs may have evolved from a green/red sensing ancestor. The ancestral proteins reconstructed in this work possess the ability also to sense blue light, which was perhaps later exploited in CBCRs with blue-light photocycles. Additionally, the ancestral photoreceptors most likely already had the ability to bind BV, which could have enabled the evolution of several extant CBCR groups that utilize BV in their photocycle and are hence able to perceive different wavelengths. The evolution of two-color sensing in the LCA of CBCR GAF domains probably made it easier to further tinker with the exact wavelengths of the *15Z* and *15E* photo-states through changes affecting the local environment and pKa of the chromophore. Our characterization of sequences representative of the first CBCR is a first step in understanding how this tinkering occurred in the colorful history of CBCR proteins.

Supplemental Figures

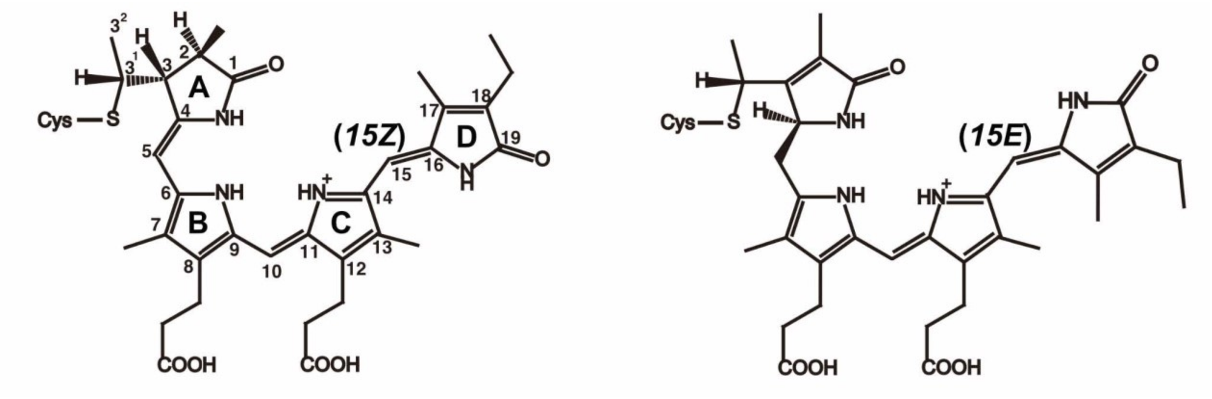
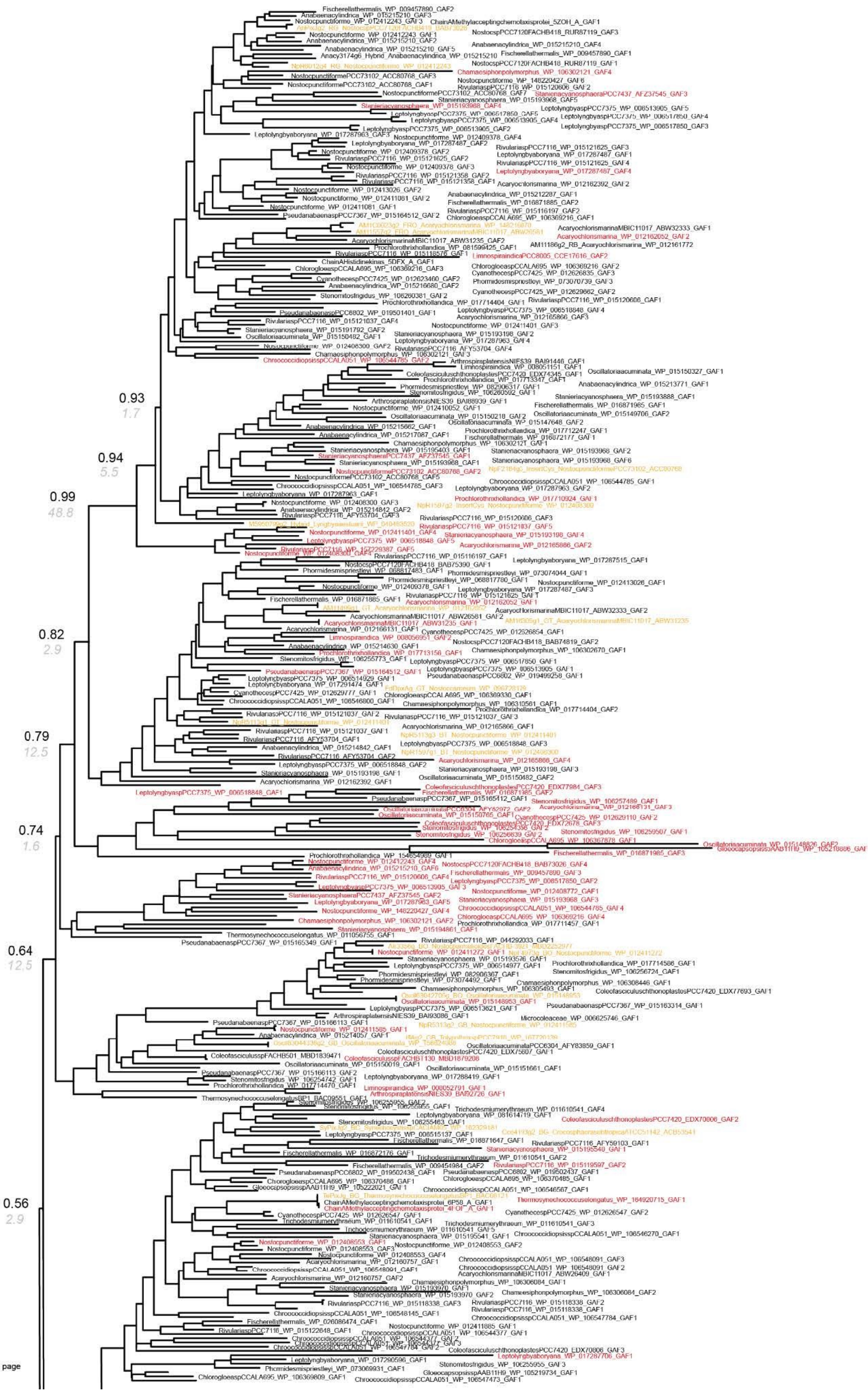


Fig. S1| Chemical structures of the phycocyanobilin chromophore bound to phytochrome/CBCR proteins. The phycocyanobilin chromophore is anchored to the "first" Cys residue via a thioether linkage at the C3^1 position of ring A. Upon absorption of a light photon, the double bond between the C and D ring isomerizes between the *15Z* configuration which is usually the dark-stable state, and the *15E* configuration which is usually a metastable photoproduct decaying thermally to the *15Z* state.



continues on next page

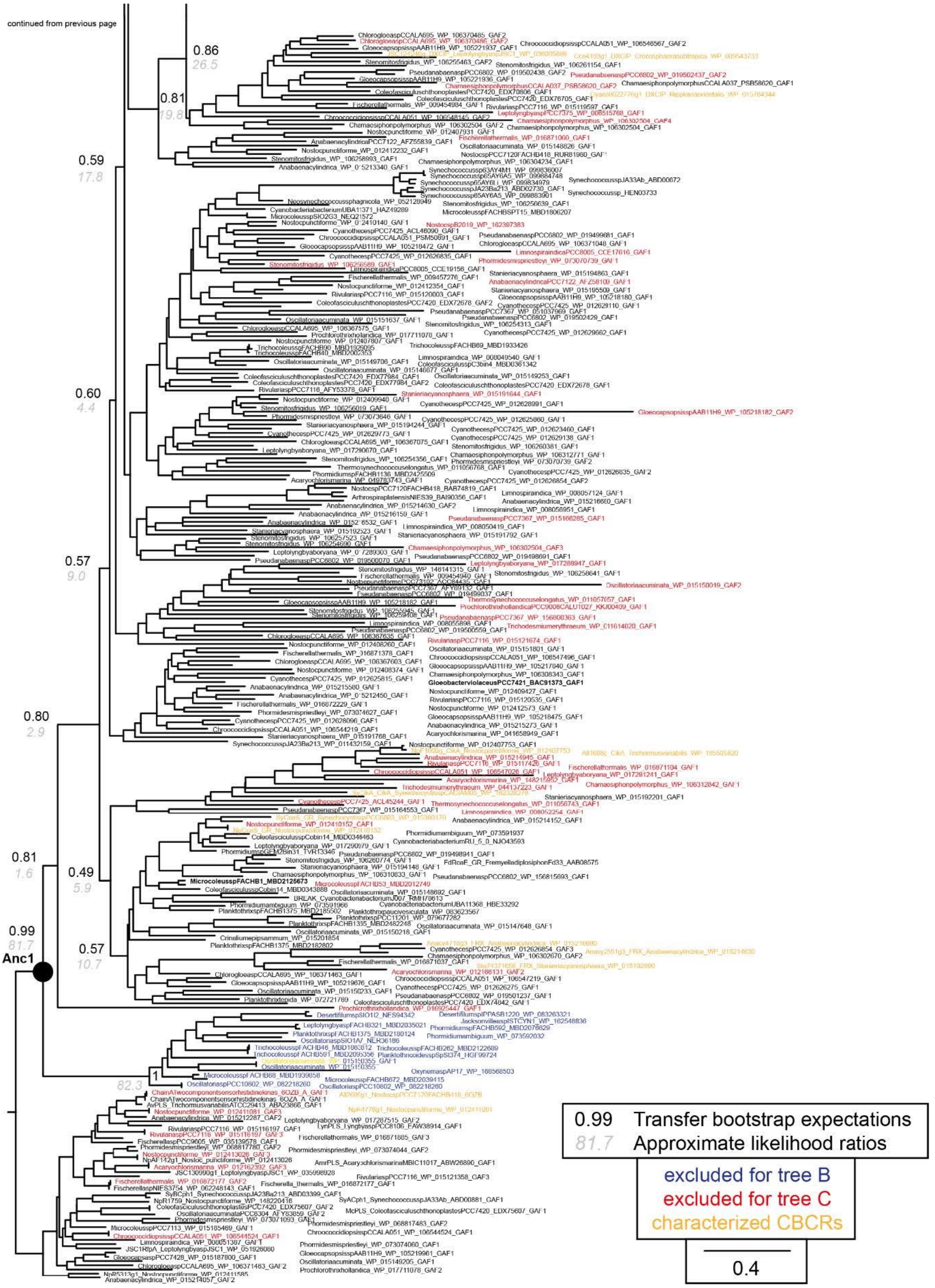


Fig. S2|. Complete phylogenetic tree of CBCR GAF domains (Tree A). The full phylogeny (of data shown in Fig. 1a) on which the ancestral CBCR GAF domain (Anc1) was reconstructed is displayed with transfer bootstrap expectations (100 replicates, black) and approximate likelihood ratios (*in gray in italic*) at critical nodes. GAF domain sequences that were removed from the corresponding alignment for the two additional trees are indicated in blue and red, respectively. CBCRs that have already been characterized biochemically are colored in orange. The scale bar represents 0.4 average substitutions per site. The tree was rooted using cyanobacterial knotless phytochromes' GAF domains as the outgroup. The complete trees for Anc2 and Anc3 can be found online in the Supplemental File 1 and 2.

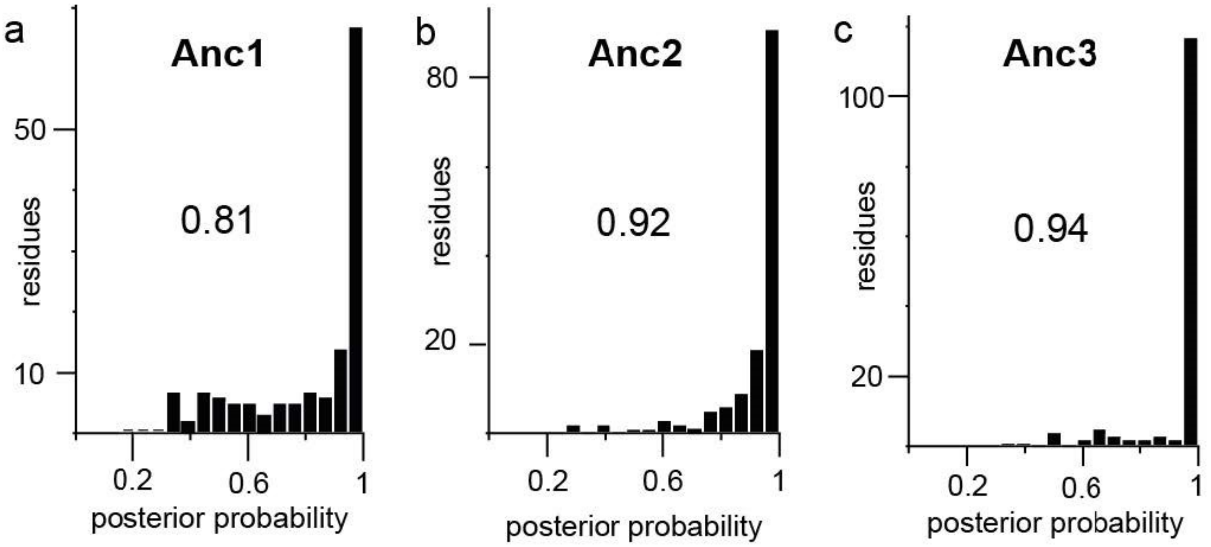


Fig. S3| Posterior probabilities of the ancestral CBCR GAF domain reconstructions. a-c, Histograms of the posterior probabilities per site with 20 bin categories and the mean.

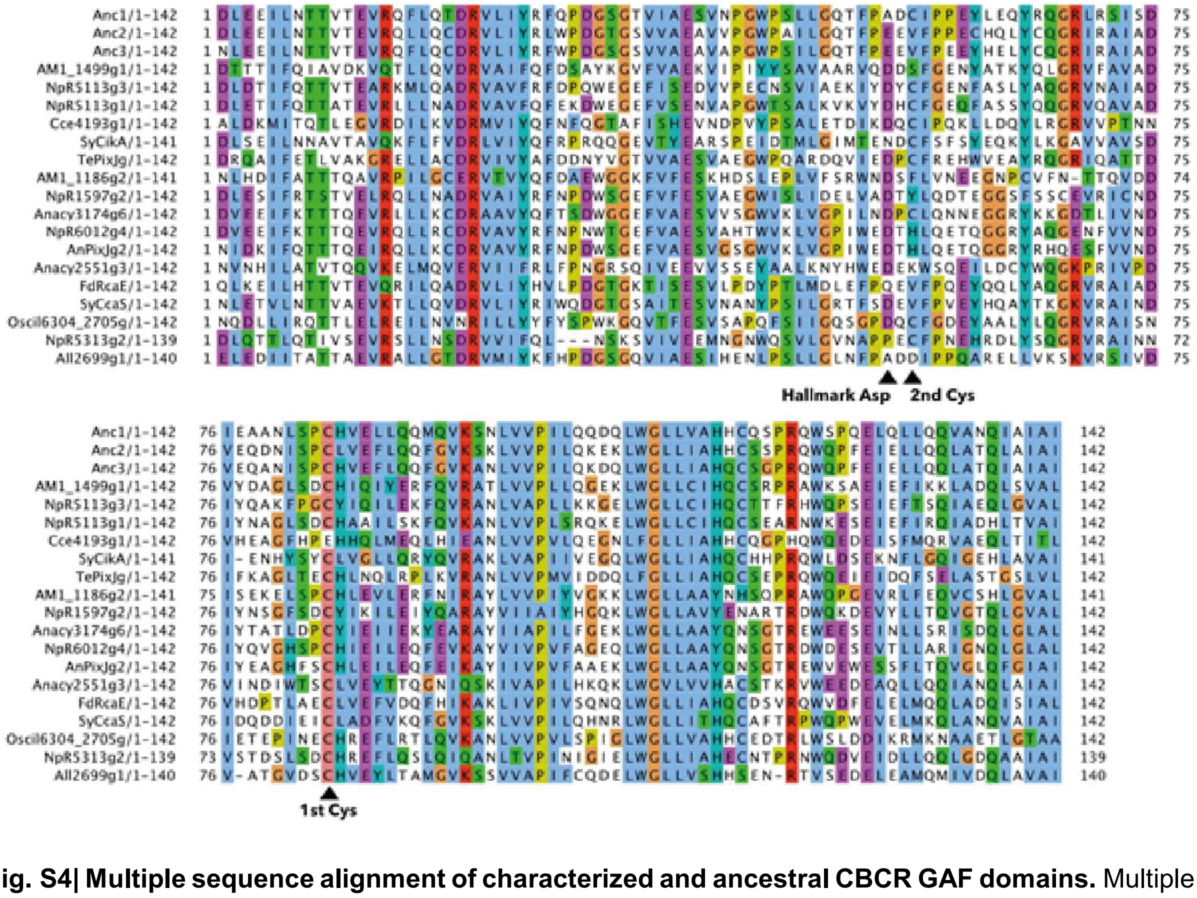


Fig. S4| Multiple sequence alignment of characterized and ancestral CBCR GAF domains. Multiple sequence alignment of exemplary GAF domain sequences which were used for the phylogenetic analysis and subsequent ancestral sequence reconstruction. Amino acids are shaded based on the characteristics of their side chain using the ClustalX-style coloring scheme. Note that *All2699g1* is the GAF domain of a PAS-less phytochrome. The amino acid positions of the chromophore-anchoring Cys (1st Cys), the other Cys binding to C10 of the chromophore for short-wavelength absorption (2nd Cys), and the Asp residue forming the hydrogen bond network with the nitrogen atoms of the chromophore (Hallmark Asp) in extant proteins are highlighted with black triangles. See Supplemental File 3 online for the complete alignments.

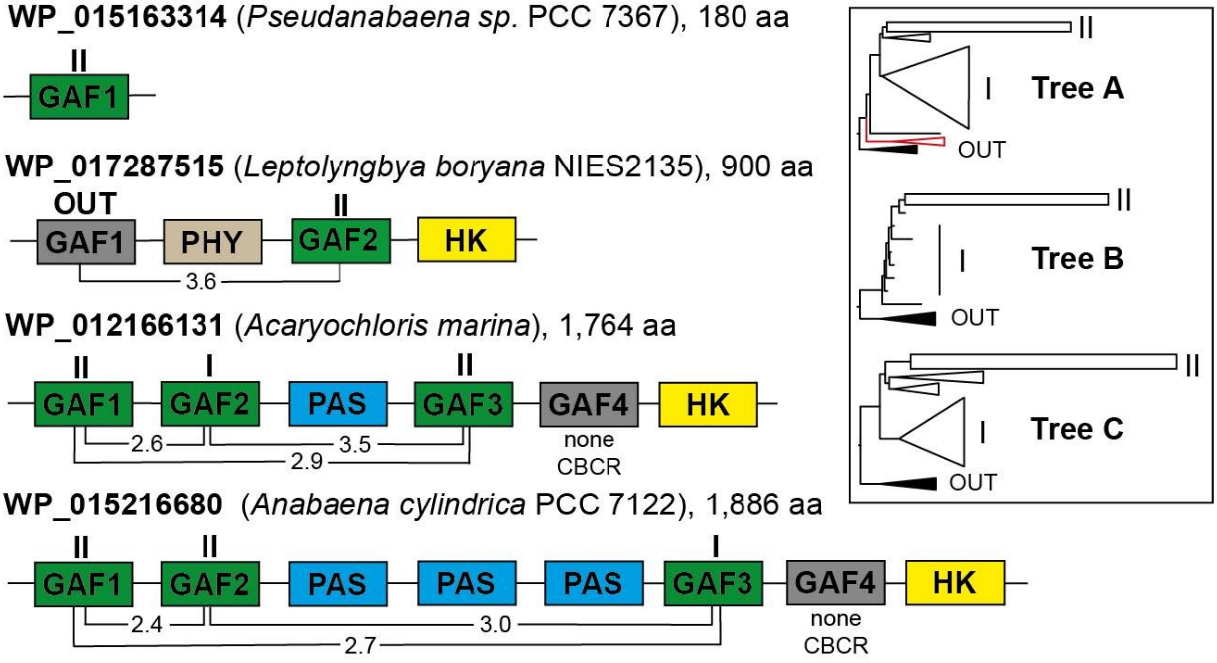

WP_015163314 (Pseudanabaena sp. PCC 7367), 180 aa
II
GAF1
WP_017287515 (Leptolyngbya boryana NIES2135), 900 aa
OUT
II
GAF1
PHY
GAF2
HK
3.6
WP_012166131 (Acaryochloris marina), 1,764 aa
II
I
II
GAF1
GAF2
PAS
GAF3
GAF4
HK
2.6
3.5
2.9
none
CBCR
WP_015216680 (Anabaena cylindrica PCC 7122), 1,886 aa
II
II
I
GAF1
GAF2
PAS
PAS
PAS
GAF3
GAF4
HK
2.4
3.0
2.7
none
CBCR
II
I
Tree A
OUT
II
I
Tree B
OUT
II
I
Tree C
OUT

Fig. S5| Various GAF domains can be found in the same polypeptide of a variety of cyanobacterial proteins. Roman numerals indicate the position of the GAF domain sequences on the phylogenetic trees (insert) as either early branching (I) or late branching CBCR GAF domains (II). OUT, outgroup of cyanobacterial knotless phytochromes' GAF domains. The evolutionary distances between domains (in branch lengths relating to Tree A in Arabic numerals) suggest early domain duplications or frequent horizontal transfer events. Serially numbered GAF, PAS, PHY, and Histidine kinase/HATPase domains (HK) are shown following the coloring scheme of Fig 1.

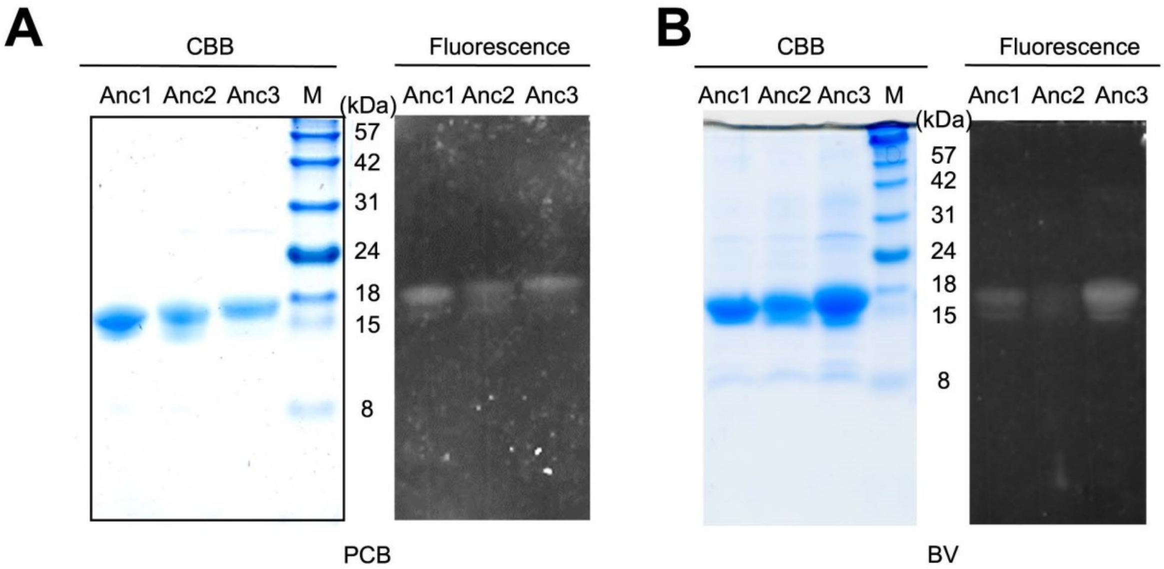


Fig. S6| SDS-PAGE of the purified ancestral CBCR proteins detected by Coomassie Brilliant Blue staining and Zn^{2+}-enhanced fluorescence imaging. **A+B**, Anc1-3 are covalently attached to the (**A**) PCB or (**B**) BV chromophore. Proteins were purified from *E. coli* and analyzed by SDS-PAGE followed by zinc acetate-enhanced fluorescence (Fluorescence) and Coomassie Brilliant Blue G-250 staining (CBB).

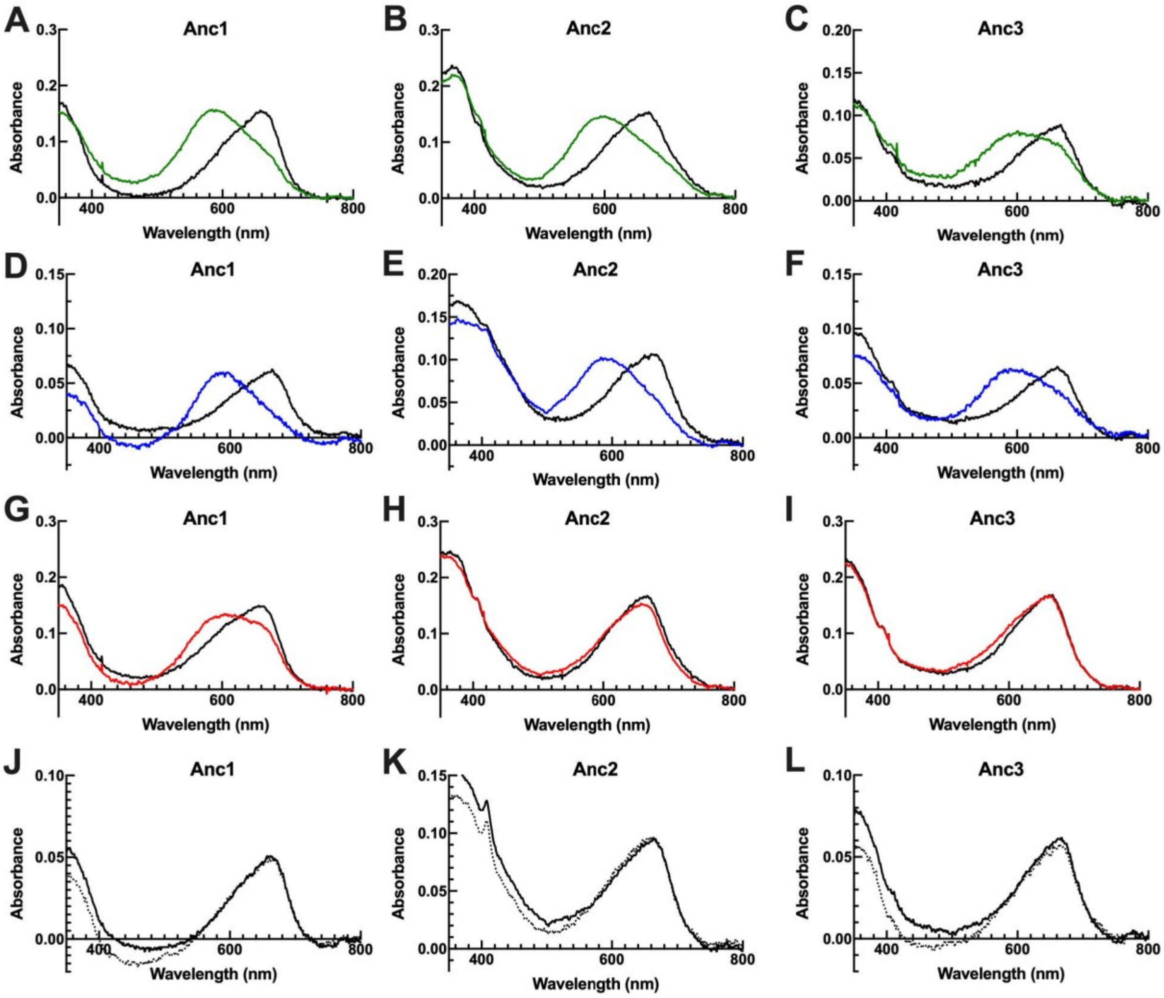


Fig. S7| Acid denaturation of the ancestral proteins with PCB chromophore. **A-C,** Absorption spectra of Anc1-3 after green irradiation for one minute, followed by denaturation by mixing with 10 M urea solution; pH 2.0 in 1:4 ratio (green lines, *15E* form) and after one minute of white-light illumination for the *15Z* form (black lines). **D-F,** Absorption spectra of *15E* (blue lines) and *15Z* (black lines) states of Anc1-3 after one minute of blue irradiation and denatured by acid urea, followed by one minute of white illumination. **G-I,** Absorption spectra of acid denatured Anc1-3 after one minute of red irradiation (red lines, *15Z* form) and after one minute of white illumination (black lines). **J-L,** Absorption spectra of acid denatured Anc1-3 after overnight incubation in the dark (dashed lines) and illumination with white light (solid lines). All experiments were performed at room temperature.

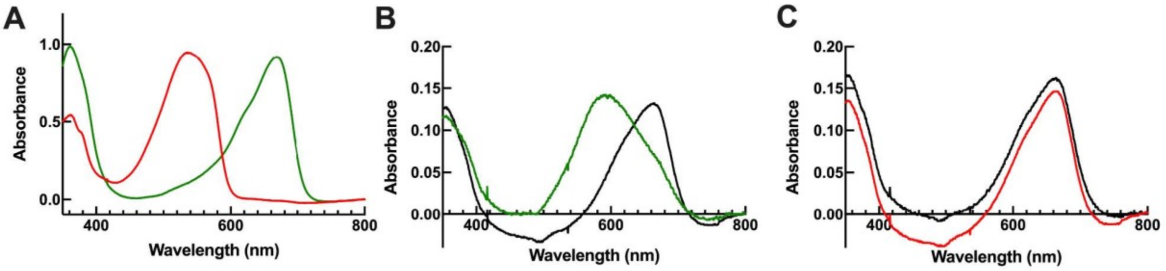


Fig. S8| Absorption and acid denaturation spectra of the early branching extant CBCR GAF domain of *Microcoleus* sp. FACHB-1 with short evolutionary distance to the reconstructed ancestral CBCR GAF domains. A, Absorption spectra after one minute of green irradiation (green line) and red irradiation (red line) of the CBCR GAF domain of *Microcoleus* sp. FACHB-1 MBD2125673. **B,** Absorption spectra after irradiation with green for one minute followed by denaturation (green line) and after one minute of white illumination (black line). **C,** Absorption spectra after irradiation with red for one minute followed by denaturation (red line) and after one minute of white illumination (black line).

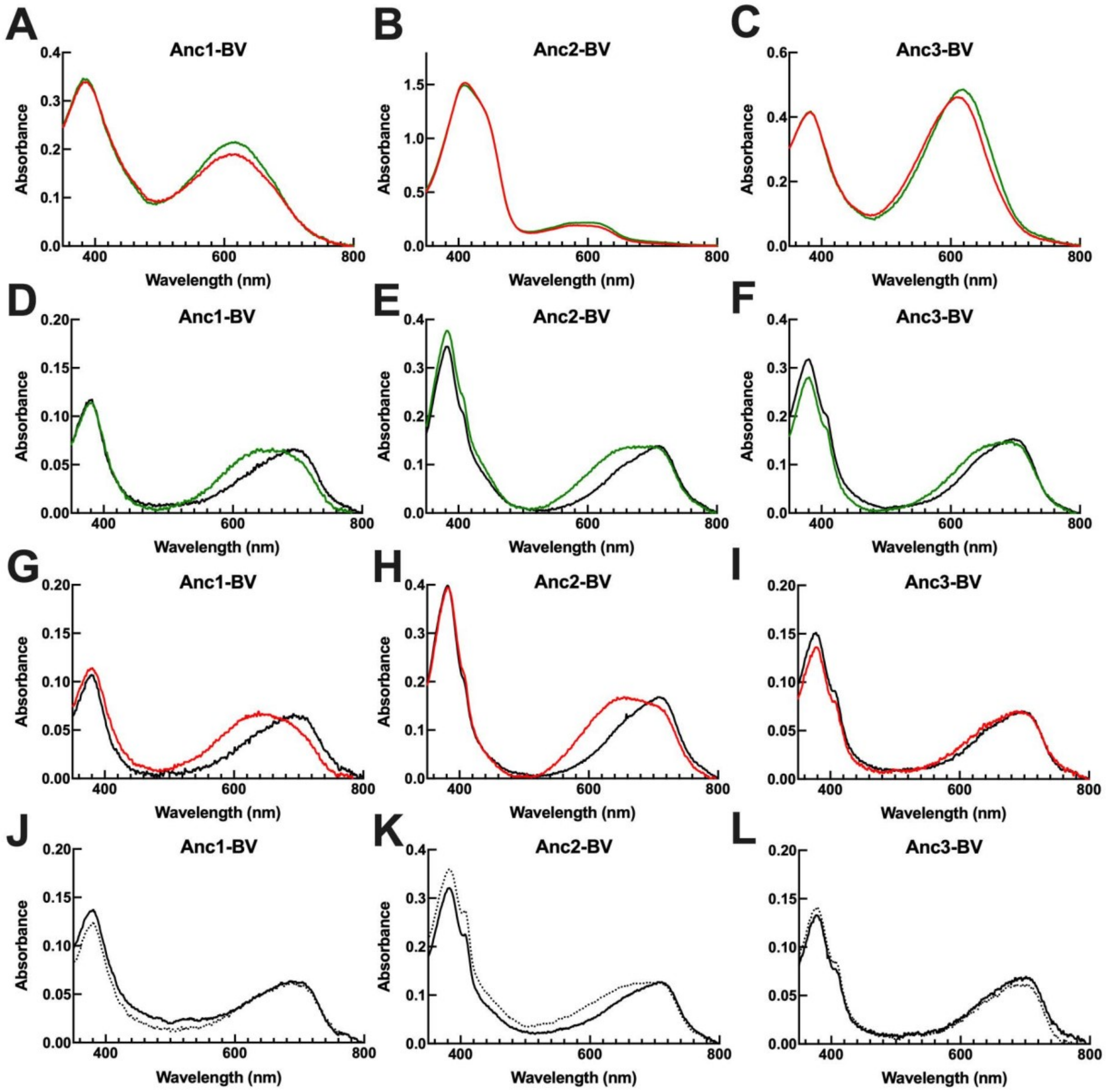


Fig. S9| Absorption and acid denaturation spectra of the ancestral proteins with BV chromophore. A-C, Absorption spectra after one minute of green irradiation (green lines) and red irradiation (red lines) of Anc1-3. **D-F,** Absorption spectra of Anc1-3 after irradiation with green for one minute followed by denaturation (green lines) and after one minute of white illumination (black lines). **G-I,** Absorption spectra of acid denatured Anc1-3 after one minute of red irradiation (red lines), followed by white illumination (black lines). **J-L,** Absorption spectra of acid denatured Anc1-3 after overnight incubation in the dark (solid lines), and after one minute of illumination with white light (dashed lines). All experiments were performed at room temperature.

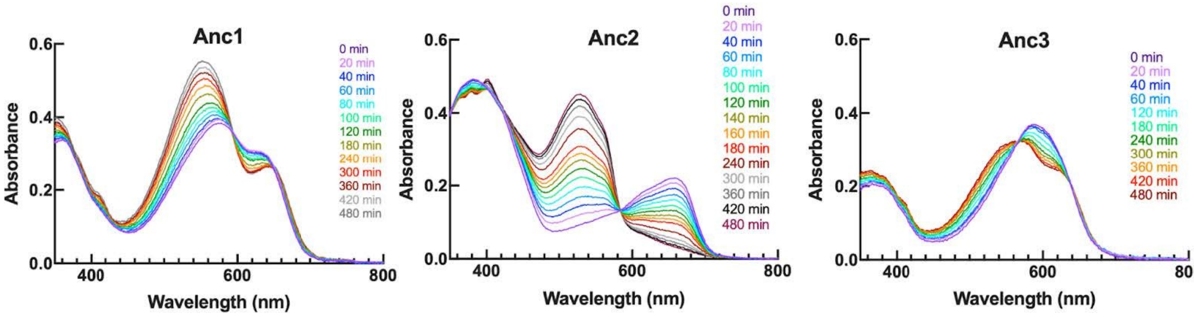


Fig. S10| Thermal reversion of the ancestral CBCR proteins. Samples were irradiated with green light for one minute to achieve the *15E* form of the protein, followed by incubation at room temperature in the dark. The absorption spectra were then recorded every 20 minutes for the first two hours for Anc1 and Anc2, and then every hour until the eighth hour. For Anc3, the absorption spectra were recorded every 20 minutes for the first hour only, due to slower thermal reversion in comparison to Anc1 and Anc2, and then every hour until the eighth hour.

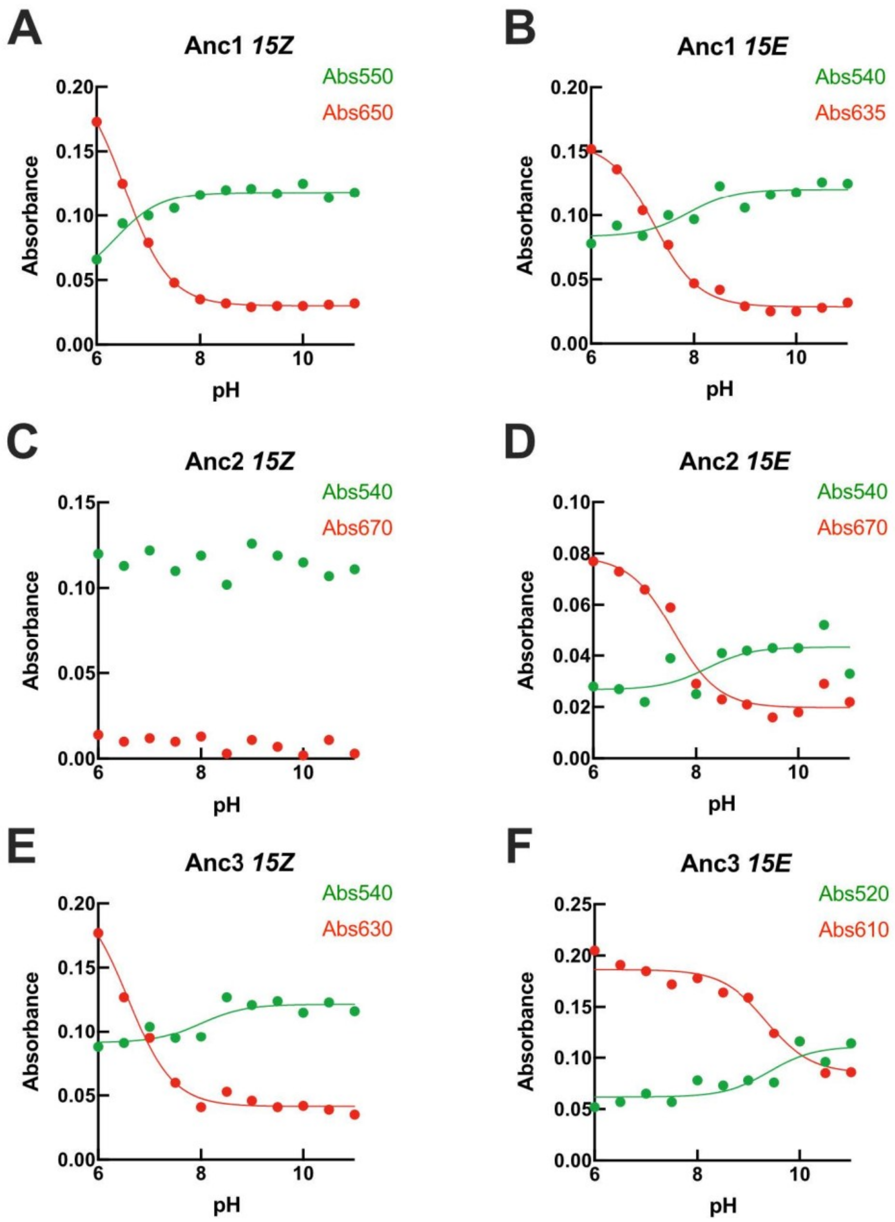


Fig. S11| pKa estimations of the ancestral CBCR wild-type proteins. A-F, pH-dependent absorption changes at indicated wavelengths of Anc1-3 with the configuration of *15Z* (**A, C, E**) or *15E* (**B, D, F**). The monitored wavelengths were selected to give the largest changes upon pH transition. Curves were fitted with one titrating group of the Henderson-Hasselbalch equation to estimate pKa.

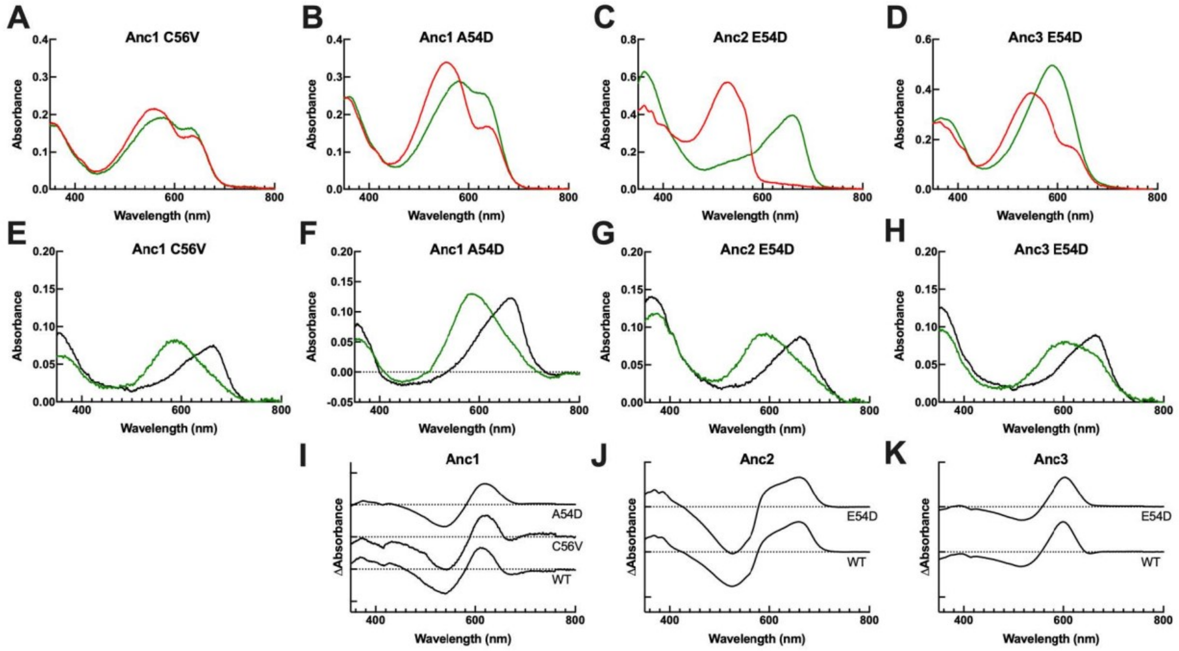


Fig. S12| Absorption and acid denaturation spectra of site-directed mutants of the ancestral CBCR proteins. A–D, Absorption spectra after one minute of green irradiation (green lines) and red irradiation (red lines). **E–H,** Absorption spectra after one minute of irradiation with green followed by denaturation (*15E*, green lines) and after one minute of white illumination (*15Z*, black lines). **I–K,** Normalized difference spectra obtained by subtracting the absorption spectra of the ^{15Z}Pg from that of the ^{15E}Pr form of Anc1-3 and their variants. Difference spectra were normalized on the red photoproduct peak and are vertically shifted for clarity. All experiments were performed at room temperature.

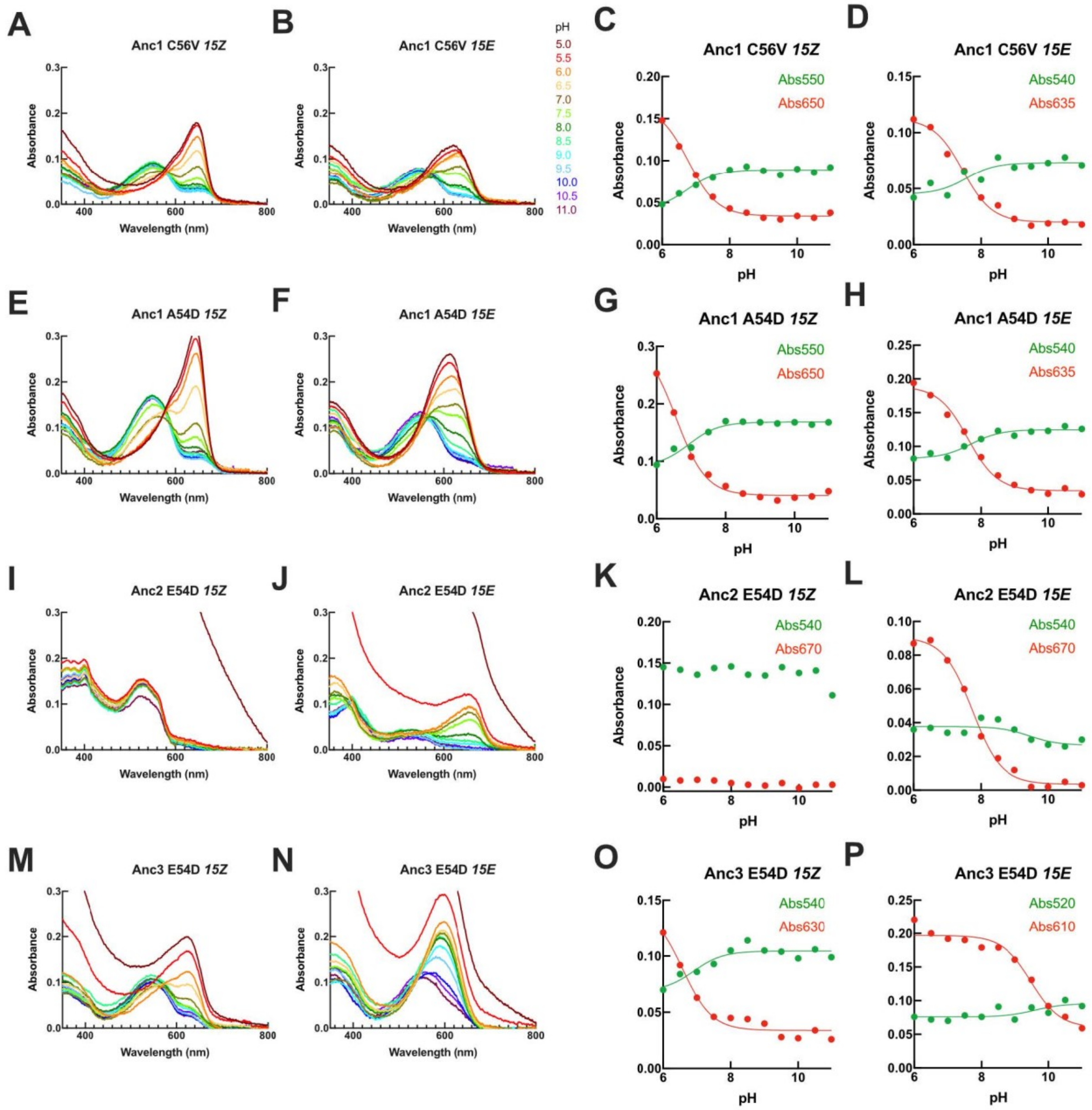


Fig. S13| pH titration analysis of site-directed mutants of the ancestral CBCR proteins. A-P, pH-dependent absorbance spectra of Anc1 C56V (**A, B**), Anc1 A54D (**E, F**), Anc2 E54D (**I, J**), and Anc3 E54D (**M, N**) with the configuration of *15Z* (**A, E, I, M**) or *15E* (**B, F, J, N**). pH-dependent absorption changes of the selected wavelengths of Anc1 C56V (**C, D**), Anc1 A54D (**G, H**), Anc2 E54D (**K, L**), and Anc3 E54D (**O, P**) with the configuration of *15Z* (**C, G, K, O**) or *15E* (**D, H, L, P**). Curves were fitted with one titrating group of the Henderson-Hasselbach equation to estimate pKa.

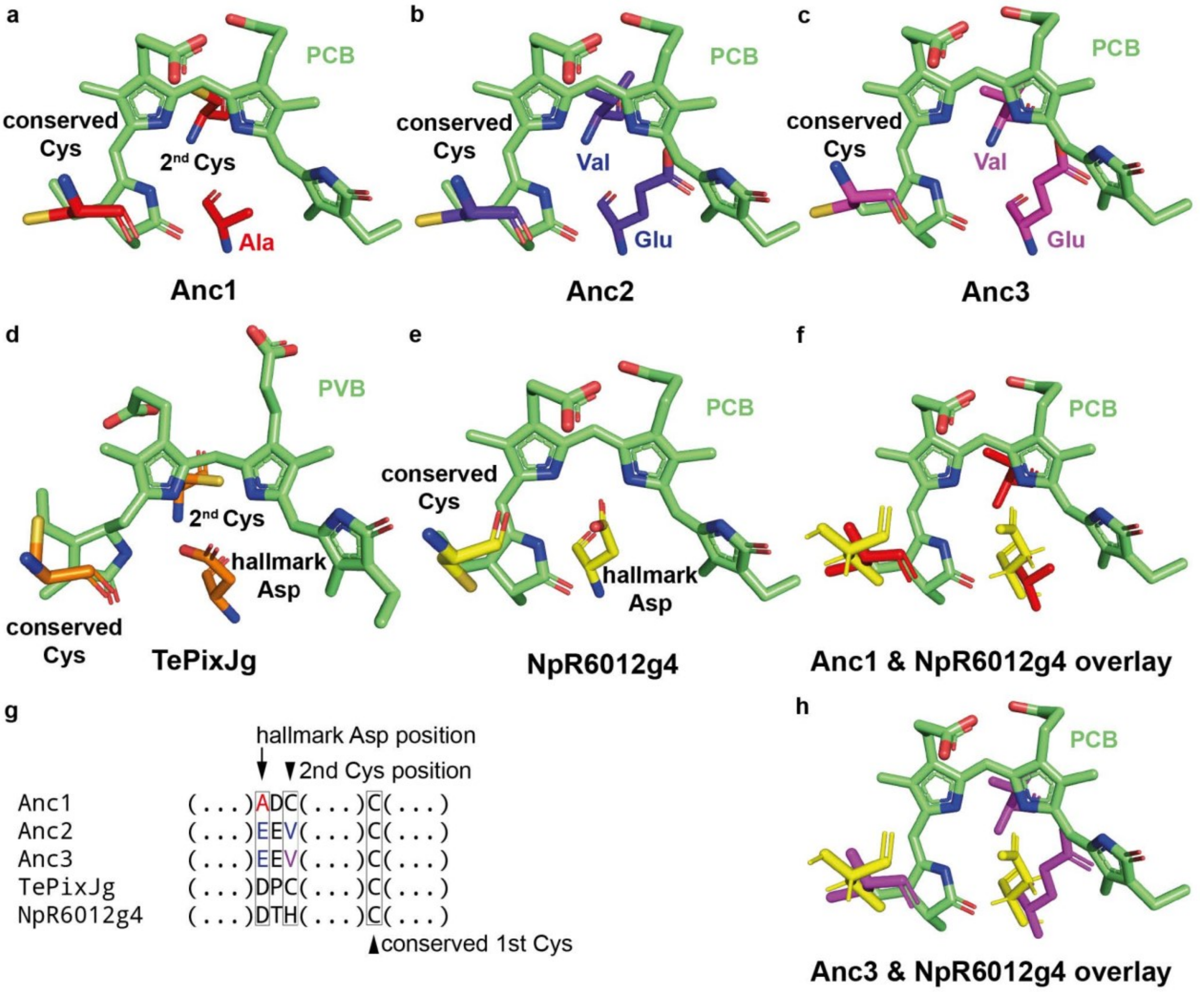


Fig. S14| Predicted structures of ancestral CBCR GAF domains and comparison with extant CBCR GAF domain structures. a–c, Structures of the important residues of (**a**) Anc1, (**b**) Anc2, and (**c**) Anc3 proteins predicted using AlphaFold2, and alignment with the PCB chromophore taken from the *NpR6012g4* structure. **d–e**, Crystal structures of *TePixJg* (PDB ID: 4GLQ) and *NpR6012g4* (PDB ID: 6BHN) with bound chromophore PVB and PCB, respectively. **f,** Overlay of the structures of Anc1 and *NpR6012g4*. **g,** Alignment of important sites of the Anc proteins with *TePixJg* and *NpR6012g4***. h,** Overlay of the structures of Anc3 and *NpR6012g4*.

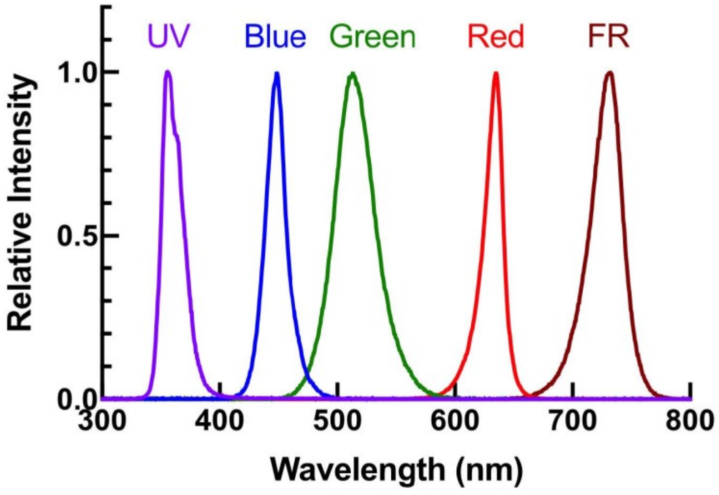


Fig. S15| Emission spectra of the used light sources. Light sources used for activation of photo-states of Anc1-3 with λ_{max} = 355 nm for UV light, 448 nm for blue light, 514 nm for green light, 635 nm for red light, and 731 nm for far-red light.

Methods

Phylogenetics and ancestral sequence reconstruction

Amino acid sequences of cyanobacterial proteins containing GAF domains were gathered using protein–protein BLAST (non-redundant protein sequences (nr) database) and a CBCR protein as a query[60]. Models (XM/XP) and uncultured/environmental sample sequences were excluded from the search. Protein sequences were manually selected to represent all large groups of the whole known cyanobacterial species phylogeny based on recently published data[61]. Sequences that were annotated to multiple species as well as incomplete sequences were excluded. Conserved domains of each sequence were identified with the HMMER web server using the Pfam database[62]. GAF domain sequences were aligned with MUSCLE 3.8[63], and the alignment was manually cropped to remove gaps by deleting lineage-specific inserts[64]. The cropped alignment was used to infer an initial ML phylogeny using RAxML[65] in the PROTGAMMAAUTO mode resulting in the LG likelihood model with fixed base frequencies. The resulting tree was rooted using GAF domain sequences of cyanobacterial proteins lacking the PAS domain but containing a PHY domain as an outgroup (cyanobacterial knotless phytochromes)[66]. The last common ancestor of all CBCR GAF domains (Anc1) was reconstructed at the internal node indicated in Fig. 1a on Tree A using the CodeML package of PAML[67] with the LG substitution model and 16 gamma categories. Due to the suspicious long branch of the 19 first branching sequences, an alternative tree (Tree B) was inferred by the deletion of these sequences from the corresponding alignment. An alternative ancestor (Anc2) was equivalently reconstructed on Tree B. For the third ancestral sequence (Anc3), Tree C was inferred after deleting all domains with particular long branches or poorly aligned sequences from the alignment. The robustness of each topology was tested by running 100 non-parametric bootstraps and calculating the transfer bootstrap estimates (TBE) for internal nodes using the BOOSTER web tool[68]. Additionally, approximate likelihood ratios were calculated with PhyML[69]. The consensus neighbor and output domains of each group on the trees were determined manually and mapped next to the topologies (Fig. 1).

Plasmid construction

Codon-optimized sequences for *E. coli* encoding the ancestral CBCR GAF domains of Anc1, Anc2, and Anc3, and *Microcoleus* sp. FACHB1 MBD2125673 WP_190776511.1 (Tab. S1, online) were obtained from Twist Bioscience (San Francisco, California, USA) or Eurofins Genomics (Ebersberg, Germany). The synthesized gene fragments were amplified by PCR and subcloned into a pET28V vector containing an N-terminal, TEV-cleavable 6×His tag via assembly cloning (AQUA cloning)[70]. Utilized oligonucleotides are provided online in Tab. S2. Sequences of the constructs were confirmed by Sanger sequencing. The PCB chromophore biosynthesis plasmid pTDho1pcyA was a kind gift from Prof. Nicole Frankenberg-Dinkel (University of Kaiserslautern)[71]. The N-terminal 6xHis tag of PcyA was removed via AQUA cloning using the primers pTDho1pcyA-1F/-2R to obtain pTDho1pcyA-HisTag. For the construction of the BV-producing plasmid, the *pcyA* gene was deleted via AQUA cloning using the primers pTDho1pcyA-3bF/-4bR to obtain the pTDho1 plasmid.

Protein production and purification

The *E. coli* strain BL21(DE3) was co-transformed with one of the pET28V plasmids harboring the gene for the target CBCR GAF domains, and either the PCB-producing pTDho1pcyA-HisTag plasmid or the BV-producing pTDho1 plasmid. The cultures were induced with 0.1 mM isopropyl-β-d-thiogalactopyranosid and grown overnight at 18 °C in LB medium with appropriate antibiotics. The cells were harvested and disrupted three times using a French cell press (50 ml, Aminco French Pressure Cell Press) at 20,000 psi in 50 mM HEPES·NaOH, pH 7.5; 300 mM NaCl, 10% (w/v) glycerol, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP), and 30 mM imidazole. The His-tagged proteins were purified by affinity chromatography with nickel affinity columns (HisTrap 1 ml; Cytiva) using the Äkta pure system (GE Healthcare UK Ltd.) from approximately 35 ml of extract. The column was washed with 10 ml of 50 mM HEPES·NaOH, pH 7.5; 300 mM NaCl, 10% (w/v) glycerol, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP), and 30 mM imidazole at a flow rate of 1 ml/min after application of the sample. Elution was carried out at a flow rate of 1 ml/min with all solutions maintained at 4 °C at a linear imidazole concentration gradient from 30 to 530 mM.

SDS-PAGE and fluorescence detection of the bound bilin chromophore

To check the purity of the protein samples, they were first denatured using 62.5 mM Tris-HCl, pH 6.8; 11.25% (w/v) glycerol, 4% SDS, 10 mM DTT, and 0.0125% (w/v) bromo-phenol blue and incubated at 95 °C for 5 min. They were separated by SDS polyacrylamide gel electrophoresis using a 16% Tris-Tricine acrylamide gel[72]. The gel was then incubated in 2 mM zinc acetate solution for 15 min and fluorescence signals were imaged using a Fusion SL (Peqlab) with an F595 Y3 filter. The gel was further stained with Coomassie G-250.

Spectroscopy and pH titration analysis

To measure the absorption spectra, the purified proteins were dialyzed in 50 mM HEPES·NaOH, pH 7.5; 300 mM NaCl, 10% (w/v) glycerol, 0.5 mM TCEP using desalting columns (HiTrap 5 ml; Cytiva), followed by irradiation with a specific wavelength for around one minute each at room temperature. The absorption spectra were acquired using a UV-2450 spectrophotometer (Shimadzu) in the dark. Thermal reversion was achieved by incubating the samples in the dark overnight at room temperature. To acquire the absorption spectra of the acid-denatured proteins, 140 µl of the protein sample was mixed with 560 µl of 10 M urea (pH 2.0) by pipetting, followed by immediate measurement of absorbance spectra.

For pH titration, the purified protein was dialyzed in 10 mM HEPES·NaOH, pH 7.5; 300 mM NaCl, 0.5 mM TCEP using desalting columns (HiTrap 5 ml; Cytiva). 560 µl of the protein was converted to either *15E* or *15Z* photo-state by irradiation of either blue, green, or red light for one minute or incubation in the dark overnight, followed by the addition of 140 µl of the following buffers in the dark (each 1 M): MES-NaOH for pH 5.0–6.5; HEPES-NaOH for pH 7.0–8.5; or glycine-NaOH for pH 9.0–11.0. The pH titration data were analyzed by fitting the absorbance value at a particular wavelength using non-linear regression in Prism software. The pKa values of the chromophore were determined using Henderson–Hasselbalch equations of a single titrating group[15,44].

Light sources

To irradiate purified proteins, LEDs illuminating at 355 nm for UV light (0.45 µmol photons m^{-2} s^{-1}), 448 nm for blue light (516 µmol photons m^{-2} s^{-1}), 514 nm for green light (540 µmol photons m^{-2} s^{-1}), 635 nm for red light (600 µmol photons m^{-2} s^{-1}), and 731 nm for far-red light (241 µmol photons m^{-2} s^{-1}) were used (Fig. S15).

AlphaFold2 structure predictions

AlphaFold2 structural predictions of the ancestral CBCR GAF domains (Anc1-3) were generated utilizing the ColabFold server on 10/18/2022 with default settings[73]. Structures were aligned to the crystal structure of the chromophore-bound *NpR6012g4* (PDB ID: 6BHN)[74] and *TePixJg* (PDB ID: 4GLQ)[46]. Data were visualized with the Pymol Molecular Graphics System v2.4.0 (Schrödinger, LLC; New York). Hallmark residues for the interaction with the chromophore in Anc1-3 were displayed in Fig. S14.

References

1. Möglich, A., Yang, X., Ayers, R.A., & Moffat, K. (2010) Structure and function of plant photoreceptors. *Annual Review of Plant Biology*, **61**, 21–47.

2. Rockwell, N.C., & Lagarias, J.C. (2010) A brief history of phytochromes. *ChemPhysChem,* **11**, 1172–1180.

3. Anders, K., & Essen, L.O. (2015) The family of phytochrome-like photoreceptors: diverse, complex and multi-colored, but very useful. *Current Opinion in Structural Biology*, **35**, 7–16.

4. Song, C., *et al.* (2014) The D-ring, not the A-ring, rotates in *Synechococcus* OS-B's phytochrome. *Journal of Biological Chemistry*, **289**, 2552–2562.

5. Klose, C., Nagy, F., & Schäfer, E. (2020) Thermal reversion of plant phytochromes. *Molecular Plant*, **13**, 386–397.

6. Jung, J.H., *et al.* (2016) Phytochromes function as thermosensors in *Arabidopsis*. *Science*, **354**, 886–889.

7. Rockwell, N.C., Martin, S.S., & Lagarias, J.C. (2012) Red/green cyanobacteriochromes: sensors of color and power. *Biochemistry*, **51**, 9667–9677.

8. Hasegawa, M., *et al.* (2018) Molecular characterization of DXCF cyanobacteriochromes from the cyanobacterium *Acaryochloris marina* identifies a blue-light power sensor. *Journal of Biological Chemistry*, **293**, 1713–1727.

9. Fushimi, K., *et al.* (2016) Cyanobacteriochrome photoreceptors lacking the canonical Cys residue. *Biochemistry*, **55**, 6981–6995.

10. Fushimi, K., Enomoto, G., Ikeuchi, M., & Narikawa, R. (2017) Distinctive properties of dark reversion kinetics between two red/green-type cyanobacteriochromes and their application in the photoregulation of cAMP synthesis. *Photochemistry and Photobiology*, **93**, 681–691.

11. Ikeuchi, M., & Ishizuka, T. (2008) Cyanobacteriochromes: a new superfamily of tetrapyrrole-binding photoreceptors in cyanobacteria. *Photochemical & Photobiological Sciences*, **7**, 1159–1167.

12. Rockwell, N.C., & Lagarias, J.C. (2020) Phytochrome evolution in 3D: deletion, duplication, and diversification. *New Phytologist*, **225**, 2283–2300.

13. Fushimi, K., & Narikawa, R. (2019) Cyanobacteriochromes: photoreceptors covering the entire UV-to-visible spectrum. *Current Opinion in Structural Biology*, **57**, 39–46.

14. Rockwell, N.C., Martin, S.S., Feoktistova, K., & Lagarias, J.C. (2011) Diverse two-cysteine photocycles in phytochromes and cyanobacteriochromes. *Proceedings of the National Academy of Sciences USA*, **108**, 11854–11859.

15. Hirose, Y., *et al.* (2013) Green/red cyanobacteriochromes regulate complementary chromatic acclimation via a protochromic photocycle. *Proceedings of the National Academy of Sciences USA*, **110**, 4974–4979.

16. Rockwell, N.C., Martin, S.S., Gulevich, A.G., & Lagarias, J.C. (2014) Conserved phenylalanine residues are required for blue-shifting of cyanobacteriochrome photoproducts. *Biochemistry*, **53**, 3118–3130.

17. Oliinyk, O.S., Chernov, K.G., & Verkhusha, V.V. (2017) Bacterial phytochromes, cyanobacteriochromes and allophycocyanins as a source of near-infrared fluorescent probes. *International Journal of Molecular Sciences*, **18**, 1691.

18. Blain-Hartung, M., Rockwell, N.C., & Lagarias, J.C. (2017) Light-regulated synthesis of cyclic-di-GMP by a bidomain construct of the cyanobacteriochrome Tlr0924 (SesA) without stable dimerization. *Biochemistry*, **56**, 6145–6154.

19. Wiltbank, L.B., & Kehoe, D.M. (2019) Diverse light responses of cyanobacteria mediated by phytochrome superfamily photoreceptors. *Nature Reviews Microbiology*, **17**, 37–50.

20. Hirose, Y., *et al.* (2019) Diverse chromatic acclimation processes regulating phycoerythrocyanin and rod-shaped phycobilisome in cyanobacteria*. Molecular Plant*, **12**, 715–725.

21. Sanfilippo, J.E., Garczarek, L., Partensky, F., & Kehoe, D.M. (2019) Chromatic acclimation in cyanobacteria: A diverse and widespread process for optimizing photosynthesis. *Annual Review of Microbiology*, **73**, 407–433.

22. Kehoe, D.M., & Grossman, A.R. (1996) Similarity of a chromatic adaptation sensor to phytochrome and ethylene receptors. *Science*, **273**, 1409–1412.

23. Enomoto, G., & Ikeuchi, M. (2020) Blue-/green-light-responsive cyanobacteriochromes are cell shade sensors in red-light replete niches. *Science*, **23**, 100936.

24. Conradi, F.D., Mullineaux, C.W., & Wilde, A. (2020) The role of the cyanobacterial type IV pilus machinery in finding and maintaining a favourable environment. *Life*, **10**, 252.

25. Ohkubo, S., & Miyashita, H. (2017) A niche for cyanobacteria producing chlorophyll f within a microbial mat. *ISME Journal*, **11**, 2368–2378.

26. Rahmatpour, N., *et al.* (2021) A novel thylakoid-less isolate fills a billion-year gap in the evolution of cyanobacteria. *Current Biology*, **31**, 2857-2867.e4.

27. Rockwell, N.C., Martin, S.S., & Lagarias, J.C. (2015) Identification of DXCF cyanobacteriochrome lineages with predictable photocycles. *Photochemical & Photobiological Sciences*, **14**, 929–941.

28. Hochberg, G.K.A., & Thornton, J.W. (2017) Reconstructing ancient proteins to understand the causes of structure and function. *Annual Review of Biophysics*, **46**, 247–269.

29. Ulijasz, A.T., & Vierstra, R.D. (2011) Phytochrome structure and photochemistry: recent advances toward a complete molecular picture. *Current Opinion in Plant Biology*, **14**, 498–506.

30. Hanson-Smith, V., Kolaczkowski, B., & Thornton, J.W. (2010) Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*, **27**, 1988–1999.

31. Berkelman, T.R., & Lagarias, J.C. (1986) Visualization of bilin-linked peptides and proteins in polyacrylamide gels. *Analytical Biochemistry*, **156**, 194–201.

32. Ishizuka, T., Narikawa, R., Kohchi, T., Katayama, M., & Ikeuchi, M. (2007) Cyanobacteriochrome TePixJ of *Thermosynechococcus elongatus* harbors phycoviolobilin as a chromophore. *Plant and Cell Physiology*, **48**, 1385–1390.

33. Moreno, M.V., Rockwell, N.C., Mora, M., Fisher, A.J., & Lagarias, J.C. (2020) A far-red cyanobacteriochrome lineage-specific for verdins. *Proceedings of the National Academy of Sciences USA*, **117**, 27962–27970.

34. Narikawa, R., *et al.* (2015) A biliverdin-binding cyanobacteriochrome from the chlorophyll d-bearing cyanobacterium *Acaryochloris marina*. *Science and Reports*, **5**, 7950.

35. Anders, K., *et al.* (2011) Spectroscopic and photochemical characterization of the red-light sensitive photosensory module of Cph2 from *Synechocystis* PCC 6803. *Photochemistry and Photobiology*, **87**, 160–173.

36. Wu, S.H., & Lagarias, J.C. (2000) Defining the bilin lyase domain: lessons from the extended phytochrome superfamily. *Biochemistry*, **39**, 13487–13495.

37. Watanabe, M., & Ikeuchi, M. (2013) Phycobilisome: architecture of a light-harvesting supercomplex. *Photosynthesis Research*, **116**, 265–276.

38. Osoegawa, S., *et al.* (2019) Identification of the deprotonated pyrrole nitrogen of the bilin-based photoreceptor by Raman spectroscopy with an advanced computational analysis. *Journal of Physical Chemistry B*, **123**, 3242–3247.

39. Nagae, T., *et al.* (2021) Structural basis of the protochromic green/red photocycle of the chromatic acclimation sensor RcaE. *Proceedings of the National Academy of Sciences USA*, **118**, e2024583118.

40. Burgie, E.S., & Vierstra, R.D. (2014) Phytochromes: An atomic perspective on photoactivation and signaling. *The Plant Cell*, **26**, 4568–4583.

41. Xu, Q.-Z., *et al.* (2019) MAS NMR on a red/far-red photochromic cyanobacteriochrome All2699 from *Nostoc. International Journal of Molecular Sciences*, **20**, 3656.

42. Wagner, J.R., Brunzelle, J.S., Forest, K.T., & Vierstra, R.D. (2005) A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. *Nature*, **438**, 325–331.

43. Narikawa, R., *et al.* (2013) Structures of cyanobacteriochromes from phototaxis regulators AnPixJ and TePixJ reveal general and specific photoconversion mechanism. *Proceedings of the National Academy of Sciences USA*, **110**, 918–923.

44. Sato, T., *et al.* (2019) Protochromic absorption changes in the two-cysteine photocycle of a blue/orange cyanobacteriochrome. *Journal of Biological Chemistry*, **294**, 18909–18922.

45. von Stetten, D., *et al.* (2007) Highly conserved residues Asp-197 and His-250 in Agp1 phytochrome control the proton affinity of the chromophore and Pfr formation. *Journal of Biological Chemistry*, **282**, 2116–2123.

46. Burgie, E.S., Walker, J.M., Phillips, G.N., & Vierstra, R.D. (2013) A photo-labile thioether linkage to phycoviolobilin provides the foundation for the blue/green photocycles in DXCF-cyanobacteriochromes. *Structure*, **21**, 88–97.

47. Rockwell, N.C., Martin, S.S., & Lagarias, J.C. (2017) There and back again: loss and reacquisition of two-Cys photocycles in cyanobacteriochromes. *Photochemistry and Photobiology*, **93**, 741–754.

48. Park, Y., Patton, J.E.J., Hochberg, G.K.A., & Thornton, J.W. (2020) Comment on "Ancient origins of allosteric activation in a Ser-Thr kinase." *Science*, **370**, eabc8301.

49. Grettenberger, C.L., *et al.* (2020) A phylogenetically novel cyanobacterium most closely related to *Gloeobacter*. *ISME Journal*, **14**, 2142–2152.

50. Watanabe, M., *et al.* (2014) Attachment of phycobilisomes in an antenna-photosystem I supercomplex of cyanobacteria. *Proceedings of the National Academy of Sciences USA*, **111**, 2512–2517.

51. Hammerschmidt, K., Landan, G., Tria, F.D.K., Alcorta, J., & Dagan, T. (2021) The order of trait emergence in the evolution of cyanobacterial multicellularity. *Genome Biology and Evolution*, **13**, 249.

52. Xu, X., *et al.* (2020) Structural elements regulating the photochromicity in a cyanobacteriochrome. *Proceedings of the National Academy of Sciences USA*, **117**, 2432–2440.

53. Essen, L.O., Mailliet, J., & Hughes, J. (2008) The structure of a complete phytochrome sensory module in the Pr ground state. *Proceedings of the National Academy of Sciences USA,* **105**, 14709–14714.

54. Anders, K., Daminelli-Widany, G., Mroginski, M.A., von Stetten, D., & Essen, L.O. (2013) Structure of the cyanobacterial phytochrome 2 photosensor implies a tryptophan switch for phytochrome signaling. *Journal of Biological Chemistry*, **288**, 35714–35725.

55. Velazquez Escobar, F., *et al.* (2017) Protonation-dependent structural heterogeneity in the chromophore binding site of cyanobacterial phytochrome Cph1. *Journal of Physical Chemistry B*, **121**, 47–57.

56. Song, C., *et al.* (2015) A red/green cyanobacteriochrome sustains its color despite a change in the bilin chromophore's protonation state. *Biochemistry*, **54**, 5839–5848.

57. Fischer, T., *et al.* (2020) Effect of the PHY domain on the photoisomerization step of the forward Pr - Pfr conversion of a knotless phytochrome. *Chemistry*, **26**, 17261–17266.

58. Narikawa, R., Enomoto, G., Ni Ni, W., Fushimi, K., & Ikeuchi, M. (2014) A new type of dual-Cys cyanobacteriochrome GAF domain found in cyanobacterium *Acaryochloris marina*, which has an unusual red/blue reversible photoconversion cycle. *Biochemistry*, **53**, 5051–5059.

59. Blain-Hartung, M., Rockwell, N.C., & Lagarias, J.C. (2021) Natural diversity provides a broadspectrum of cyanobacteriochrome-based diguanylate cyclases. *Plant Physiology*, **187**, 632–645.

60. Altschul, S. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

61. Moore, K.R., *et al.* (2019) An expanded ribosomal phylogeny of cyanobacteria supports a deep placement of plastids. *Frontiers in Microbiology*, **10**, 1612.

62. Potter, S.C., *et al.* (2018) HMMER web server: 2018 update. *Nucleic Acids Research*, **46**, W200–W204.

63. Madeira, F., *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, **47**, W636–W641.

64. Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

65. Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

66. Ulijasz, A.T., *et al.* (2008) Characterization of two thermostable cyanobacterial phytochromes reveals global movements in the chromophore-binding domain during photoconversion. *Journal of Biological Chemistry*, **283**, 21251–21266.

67. Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

68. Lemoine, F., *et al.* (2018) Renewing Felsenstein’s phylogenetic bootstrap in the era of big data. *Nature*, **556**, 452–456.

69. Guindon, S., *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

70. Beyer, H.M., *et al.* (2015) AQUA cloning: a versatile and simple enzyme-free cloning approach. *PLoS one*, **10**, e0137652.

71. Dammeyer, T., Bagby, S.C., Sullivan, M.B., Chisholm, S.W., & Frankenberg-Dinkel, N. (2008) Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Current Biology*, **18**, 442–448.

72. Schägger, H. (2006) Tricine–SDS-PAGE. *Nature Protocols*, **1**, 16–22.

73. Mirdita, M., *et al.* (2022) ColabFold: making protein folding accessible to all. *Nature Methods*, **19**, 679–682.

74. Lim, S., *et al.* (2018) Correlating structural and photochemical heterogeneity in cyanobacteriochrome NpR6012g4. *Proceedings of the National Academy of Sciences USA*, **115**, 4387–4392.

Fortuitously compatible protein surfaces primed allosteric control in cyanobacterial photoprotection

This work is published in:

Steube, N.[1] & Moldenhauer, M.[1], Weiland, P., Saman, D., Kilb, A., Ramírez Rojas, A.A., Garg, S.G., Schindler, D., Graumann, P.L., Benesch, J.L.P., Bange, G., Friedrich, T. & Hochberg, G.K.A. Fortuitously compatible protein surfaces primed allosteric control in cyanobacterial photoprotection. *Nature Ecology and Evolution* **7**, 756–767 (2023); https://doi.org/10.1038/s41559-023-02018-8.

[1] Niklas Steube & Marcus Moldenhauer have contributed equally to this work.

The online version of the publication is open access and contains all supplementary material, available at https://doi.org/10.1038/s41559-023-02018-8.

Source data are available at the Open Research Data Repository of the Max Planck Society (Edmond) under the https://doi.org/10.17617/3.44RHFZ. Crystallography data are available at RCSB PDB under accession number 8AG8. Sequencing data are available on NCBI Sequence Read Archive under BioProject PRJNA865569.

Own contribution:
Niklas Steube performed phylogenetic analyses and ancestral protein sequence reconstructions. He cloned, purified, and characterized extant and ancestral FRP and FRPL proteins and performed circular dichroism spectroscopy and genetic manipulation of *Pseudomonas borbori*. He further performed AlphaFold2 protein interaction prediction analyses, designed figures, and contributed in discussing all data and writing the manuscript.

Co-author's contribution:
NS, MM, TF and GKAH conceived the project and oversaw the manuscript writing. MM performed protein purification, biophysical and biochemical experiments. PW and GB performed protein crystallography and interpreted the data. AK and PLG performed epi-fluorescence microscopy and interpreted the data. D.Saman and JLPB performed native mass spectrometry and interpreted the data. AARR and D.Schindler sequenced *P. borbori* and analyzed the data. SGG inferred the phylogenetic species trees and performed the gene tree–species tree reconciliation. NS, MM, TF and GKAH interpreted all data. All authors contributed to manuscript writing and discussion.

Abstract

Highly specific interactions between proteins are a fundamental prerequisite for life, but how they evolve remains an unsolved problem. In particular, interactions between initially unrelated proteins require that they evolve matching surfaces. It is unclear whether such surface compatibilities can only be built by selection in small incremental steps, or whether they can also emerge fortuitously. Here, we used molecular phylogenetics, ancestral sequence reconstruction and biophysical characterization of resurrected proteins to retrace the evolution of an allosteric interaction between two proteins that act in the cyanobacterial photoprotection system. We show that this interaction between the orange carotenoid protein (OCP) and its unrelated regulator, the fluorescence recovery protein (FRP), evolved when a precursor of FRP was horizontally acquired by cyanobacteria. FRP's precursors could already interact with and regulate OCP even before these proteins first encountered each other in an ancestral cyanobacterium. The OCP–FRP interaction exploits an ancient dimer interface in OCP, which also predates the recruitment of FRP into the photoprotection system. Together, our work shows how evolution can fashion complex regulatory systems easily out of pre-existing components.

Introduction

Allosteric interactions between proteins are a ubiquitous form of biochemical regulation in which the active site of one protein is affected by binding of another protein to a distal site[1]. How such interactions evolve is an unsolved problem in evolutionary biochemistry. It requires that both proteins (the regulator and the target) evolve a matching interface as well as some mechanism that translates binding of the regulator to a change at the active site of the target protein. If all residues that participate in this interface and the transmission mechanism have to evolve *de novo*, building such an interaction would require several substitutions in both proteins. Because long genetic trajectories involving several substitutions in multiple proteins are very unlikely to be fixed by random genetic drift, existing interactions are usually assumed to have been built up in incremental mutational steps. Each step would add a single interacting residue and would be driven to fixation by natural selection acting directly on a function associated with the interaction[2]. However, in a few protein systems, interfaces or allosteric pathways pre-existed fortuitously in one of the two partners[3-6]. This indicates that some aspects of these interactions arose by chance, which were then exploited by other components that arose later.

It remains unclear to what extent direct selection is necessary to fashion these remaining components of an interaction, such as the interaction surface of a new regulator that exploits a pre-existing surface on its target. In principle, these features could also be entirely accidental if they initially fixed for reasons unrelated to the interaction. In all well-studied cases we cannot answer this question because both components originated from within the same genome where the target and the regulator would have always encountered each other, so selection may or may not have acted to adapt the regulator to its new target[3-6]. Whether any biologically meaningful interaction ever truly arose by chance therefore remains unknown.

Here, we address this problem by studying the evolution of an allosteric interaction in the cyanobacterial photoprotection system[7,8]. Photoactive organisms must protect themselves from high light irradiation causing photodamage. In cyanobacteria, this protection is mediated by the orange carotenoid protein (OCP)[9,10], a photoactive light intensity sensor with a carotenoid embedded symmetrically into its two domains that is able to switch conformation from an inactive orange (OCP^{O}) to an activated red state (OCP^{R}) under high light conditions[11]. Activated OCP^{R} binds to the cyanobacterial light-harvesting antenna complex, the phycobilisome, to dissipate excess phycobilisome excitation as heat[11,12]. Two OCP paralogues (OCP2 and OCPx) can detach from the phycobilisome and recover into OCP^{O} passively in the dark[11,13]. However, the most common paralogue OCP1 relies on an allosteric regulation for photo-recovery: OCP1 interacts with the fluorescence recovery protein (FRP), a small, dimeric regulator that terminates the interaction with the phycobilisome, and strongly accelerates the back-conversion of OCP^{R} into the resting orange state[14,15] (Fig. 1a). Although the likely evolution of OCP from non-photo-switchable precursors has recently been demonstrated[16], it is not yet known how FRP was recruited into the cyanobacterial photoprotection system as a new allosteric regulator.

Results

Ancestral OCPs are photo-switchable light intensity sensors

To retrace the evolutionary origins of OCP1's allosteric interaction with FRP, we first sought to understand how OCP paralogues evolved and when they gained the ability to be regulated by FRP. It has recently been shown that the first OCP probably evolved via a gene fusion event of two small proteins and that a linker addition provided photo-switchability[16]. Homologues of these single domain proteins can still be found in extant cyanobacteria, and have been termed helical carotenoid proteins (HCPs) and C-terminal domain-like homologues (CTDHs) that feature a common fold of nuclear transport factor 2 proteins (NTF2)[17]. We first inferred a maximum likelihood (ML) phylogeny of OCP proteins, using cyanobacterial CTDH sequences as the outgroup to root our tree (Fig. 1b, Extended Data Fig. 1). We further describe an alternative rooting using HCP sequences in Extended Data Fig. 2. Our phylogenetic tree is virtually identical to a recently published tree[16], with OCPx, OCP2 and OCP1 each forming well-supported monophyletic groups. OCP1 and OCP2 are sister groups, to the exclusion of all other OCPs. Two more uncharacterized clades branch between the OCPx group and OCP1 and OCP2, which could be additional OCPx or represent separate paralogues.

We used ancestral sequence reconstruction to infer the amino acid sequences of ancestral OCPs at the internal nodes of our tree and along the lineage towards FRP-regulated OCP1. We focused on three proteins from the last common ancestor (LCA) of all extant OCP (AncOCPall) to the LCA of OCP1 and OCP2 paralogues (AncOCP1&2) up to the LCA of extant OCP1 (AncOCP1), which were reconstructed with average posterior probabilities across sites between 0.92 and 0.96 (Fig. 1b, Extended Data Fig. 3a–e). We resurrected these ancestral OCP proteins heterologously in *Escherichia coli*, and purified them for *in vitro* characterization.

All ancestral OCPs are photo-switchable light intensity sensors with a bound echinenone as the favored carotenoid (Fig. 1c, Extended Data Fig. 4a–h). AncOCPall shows a moderate time constant for the OCP^R to OCP^O back-conversion of 166 ± 10 s (similar to extant OCP2[16]). The recovery constant decreases to 20 ± 1 s in AncOCP1&2 (faster than extant OCPs), but drastically increases in AncOCP1 to 314 ± 8 s (as in extant OCP1) (Fig. 1d–f, Extended Data Fig. 4i–l). Our data show that slow photo-recovery is a feature that evolved along the branch to OCP1, consistent with the theory that only OCP1 paralogues require FRP for allosterically accelerated recovery.

FRP-accelerated recovery evolved along the branch leading to OCP1

We next tested the effect of an extant FRP from *Synechocystis* sp. PCC 6803 on the recovery times of our ancestral OCPs. The two earlier ancestors are unaffected by FRP, whereas AncOCP1 is only able to rapidly recover in the presence of FRP (in molar ratios of five OCP to one FRP), which accelerates the OCP^{R} to OCP^{O} back-conversion by about 97% (similar to extant OCP1) (Fig. 1d-f, Extended Data Fig. 4m-t). As AncOCP1&2 is unaffected by FRP, the allosteric acceleration of OCP's recovery evolved after the gene duplication event that gave rise to OCP1 and OCP2 paralogues, only along the branch to OCP1.

We tested the robustness of our conclusions to statistical uncertainties in our resurrected sequences by additionally resurrecting one less likely, but still statistically plausible, alternative sequence per ancestor (see Methods for details). Biophysical characterizations of these alternative ancestral OCP proteins confirm that slow recovery and acceleration by FRP evolved along the branch leading to OCP1(Extended Data Fig. 5a–l).

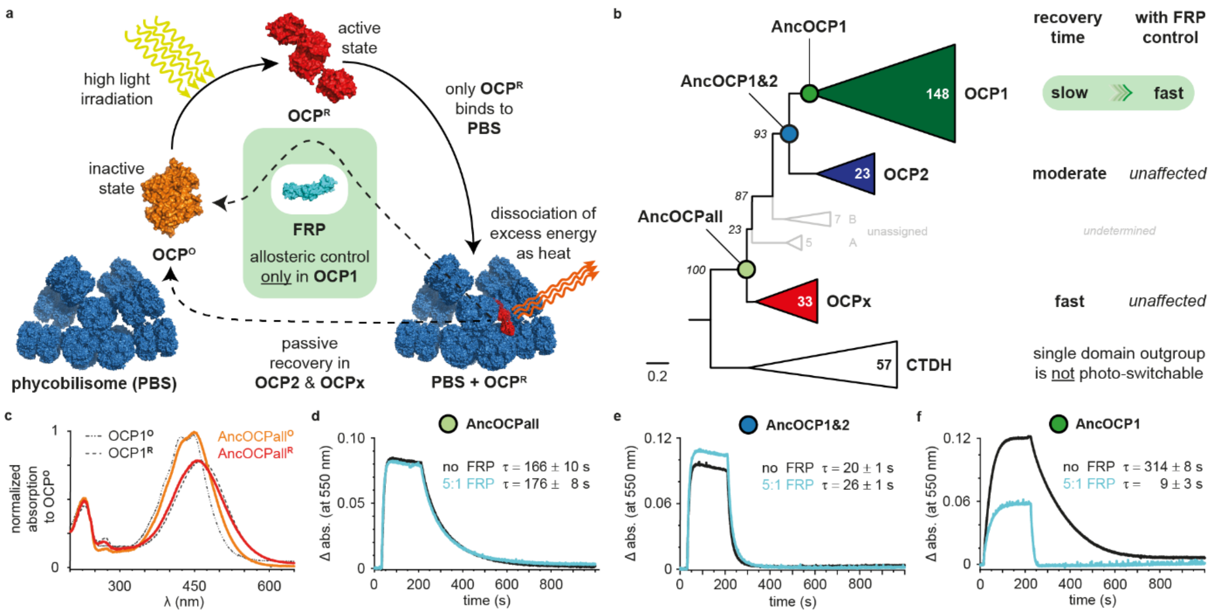


Fig. 1| Evolution of allosteric control in OCP. a, Mechanism of cyanobacteria-exclusive, OCP-mediated photoprotection involving allosteric control by FRP (cyan) in OCP1 paralogues. Structures used (PDB IDs): 7EXT[57],3MG1[58], 4JDX[25], and 7SC9[29]. **b**, Reduced ML phylogeny of OCP paralogues with relative speed of recovery from photoconversion indicated, and reconstructed ancestral proteins (Anc) of selected clades. Cyanobacterial CTDHs are the outgroup. Bold numbers count taxa of designated OCP paralogues. Italic numbers are Felsenstein bootstrap probabilities of 100 replicates. Branch-lengths represent average substitutions per site. The complete tree is shown in Extended Data Fig. 1. **c**, Ultraviolet–visible absorption spectra of inactive orange and active red state of AncOCPall in comparison with extant OCP1 from *Synechocystis* sp. PCC 6803 (SYNY3; dashed lines). **d–f**, Recovery from photoconversion of ancestral OCPs at 20 °C with (cyan) or without SYNY3 FRP (black), and respective mean recovery time constants (τ) with SD of three independent replicates: AncOCPall (**d**), AncOCP1&2 (**e**) and AncOCP1 (**f**). Representative data sets are shown for clarity.

FRP was acquired horizontally early in cyanobacterial history

We next asked when FRP first appeared in cyanobacterial genomes, relative to the gene duplication that produced FRP-regulated OCP1. To answer this, we inferred a ML species phylogeny of OCP-containing cyanobacterial strains found on our OCP tree and mapped the presence of FRP and OCP paralogues onto it (Extended Data Fig. 6). Virtually all OCP1-containing genomes also contain FRP, suggesting FRP was gained close in time to the duplication that produced OCP1. Exactly where on the species phylogeny the successive OCP duplications occurred is difficult to tell, because OCP2 and OCPx paralogues have very sporadic distributions, and the relationships within each OCP clade are only poorly resolved. Gloeobacteria, which on our and others' species phylogenies[18-21] are sister to all other cyanobacteria, only possess OCPx, whereas groups branching immediately after already have OCP1 and FRP or OCP2 or both. This suggests that the duplication that produced OCP1 and OCP2 happened relatively quickly after *Gloeobacter* spp. split off from all other cyanobacteria, and that FRP was recruited into the system around the same time.

Our next goal was to understand the origin of FRP. Homologues of FRP (termed FRP-like, FRPL) can also be found in distantly related bacteria[8,22], mainly proteobacteria and acidobacteria, suggesting an origin far beyond cyanobacteria. To test this theory, we extensively searched for FRP homologues in and outside cyanobacteria and inferred a ML phylogeny. Our tree features a highly supported split between all FRPs and all FRPLs (Fig. 2a). A small group of delta-proteobacterial FRPLs branches closest to the cyanobacterial FRP group with high statistical support (approximate likelihood-ratio test (aLRT) = 60.9, transfer bootstrap expectations (TBE) of 0.99). However, in some bootstrap runs FRPLs of other bacterial taxa with long terminal branches jump into this group, resulting in poor Felsenstein bootstrap support (FBP = 0.51), but the delta-proteobacterial FRPLs remain sister to FRP in all runs. Further FRPLs are sporadically distributed in the proteobacteria and acidobacteria, and mostly found in uncultured species (and entirely absent in model organisms). Within different groups of proteobacteria our tree becomes poorly resolved, probably owing to the short length of FRP and FRPL proteins.

We rooted the tree between acidobacteria and proteobacteria within the FRPL group as the most parsimonious root hypothesis. This root indicates a horizontal gene transfer (HGT) from an ancestral delta-proteobacterium into an ancestral cyanobacterium, and further indicates many sporadic losses of FRPL in acidobacteria and proteobacteria (Fig. 2a). A root within the FRP group would in contrast require more and less plausible HGT events: at least from cyanobacteria into only a small set of proteobacteria, then into acidobacteria and then from relatively modern acidobacteria into early proteobacteria. A root between FRPs and FRPLs would require an origin of the protein in the LCA of all bacteria[23], which would indicate losses in many large bacterial groups as well as the same temporally implausible transfer from modern acidobacteria into the LCA of proteobacteria (see Supplementary Discussion for details). As a consequence, our results indicate that FRP was most probably horizontally acquired by an ancestral cyanobacterium early in cyanobacterial history.

FRP evolved from structurally highly similar proteins

To understand the ancestral state of FRPL proteins before they were transferred into cyanobacteria, we heterologously expressed, purified and characterized the FRPL from one of the few isolated, mesophilic bacteria that feature FRPL (PbFRPL): the gamma-proteobacterium *Pseudomonas borbori*, a close relative of *P. aeruginosa*[24]. Circular dichroism spectroscopy of PbFRPL showed the typical all alpha-helical fold, previously found in FRP in solution, and native mass spectrometry confirmed the distinctive dimeric state[8,14] (Extended Data Fig. 7a–c). We solved PbFRPL's crystal structure to a resolution of 1.8 Å (Tab. 1). The N-terminal domain consists of two antiparallel alpha-helices of about 50 Å in length and features a homo-dimerization interface similar to those in FRPs with an estimated buried surface of around 675 Å^2. The C-terminal head domain, that in FRP is thought to interact with OCP1[25-27], is also present in PbFRPL, and constitutes three interlocking alpha-helices. Overall, PbFRPL and FRP from *Synechocystis* sp. PCC 6803 (Protein Data Bank (PDB) ID 4JDX[25] superpose with a root-mean-square deviation of 2.08 Å (Fig. 2b,c). PbFRPL's structural properties are therefore extremely similar to those of cyanobacterial FRP.

It is unclear what function FRPLs carry out, but it cannot be regulating OCP because genomes containing FRPL contain neither OCPs nor homologues of their N- or C-terminal domain-like proteins (HCP and CTDH, respectively). In *P. borbori*, the *frpl* gene is encoded on its single chromosome, and we did not find any OCP, HCP or CTDH homologues (Extended Data Fig. 7d).

Epi-fluorescence microscopy of PbFRPL fused to an mVenus fluorophore and expressed from a plasmid under its native promotor in *P. borbori* showed a homogeneous distribution across the whole cell during exponential growth and an additional concentration at the cell poles upon starvation with increased whole-cell integrated fluorescence by about 2.5- to 3.4-fold above wild-type increase (Extended Data Fig. 7e–g). Keeping in mind that we cannot control for protein copy number here, it is noticeable that PbFRPL localization and quantity change in response to starvation.

Our data indicate that despite their extremely similar structures, FRPLs carry out a potentially stress-related function that must be totally unrelated to OCPs and the regulation of photoprotection.

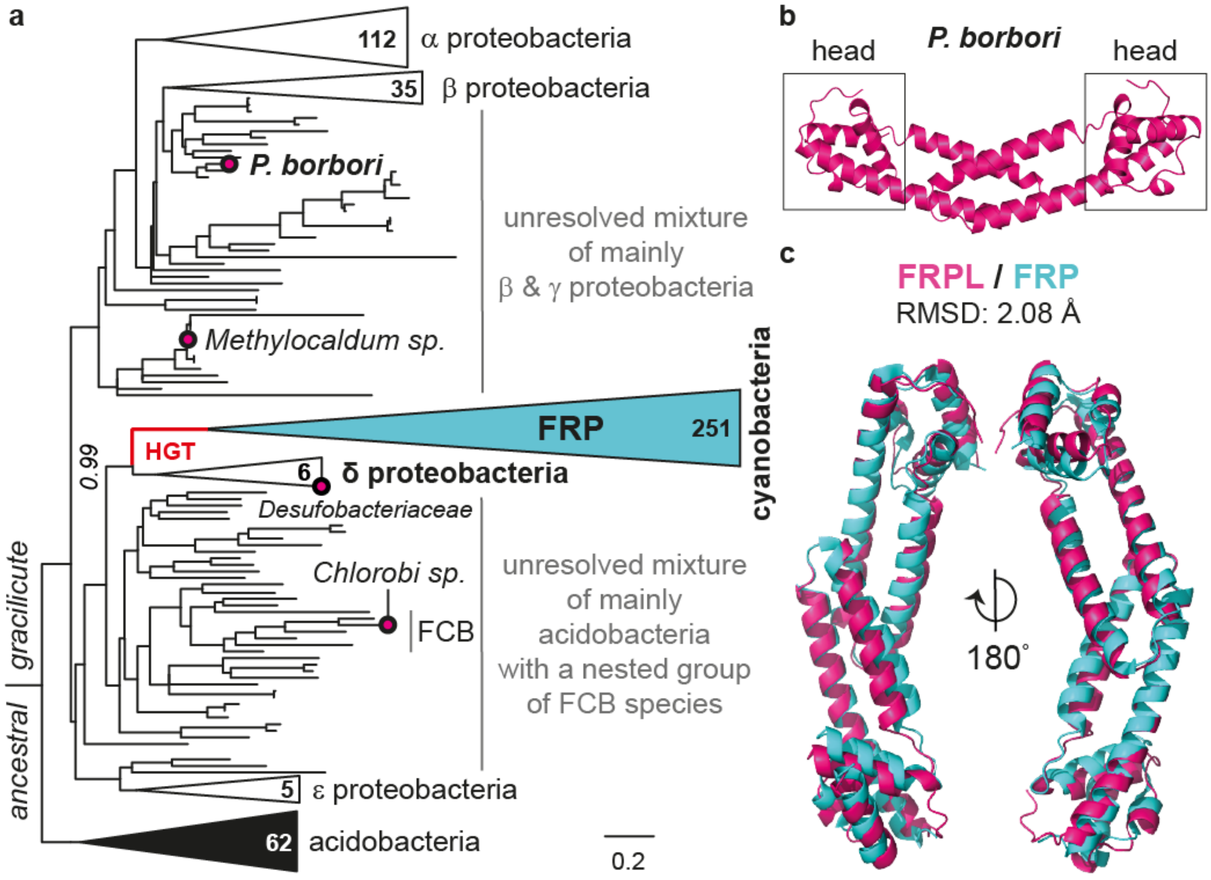


Fig. 2| FRP evolved from structurally highly similar proteins through horizontal transfer. a, Reduced ML phylogeny of cyanobacterial FRP (cyan), and homologous FRPL proteins with examined ones in this study indicated by a magenta circle and their host species' name. Bold numbers count taxa of collapsed bacterial groups. Italic number indicates TBE of 100 replicates. The tree was rooted between proteobacteria and acidobacteria, and indicates a HGT between delta-proteobacteria and cyanobacteria (red line). Branch lengths represent average substitutions per site. The complete tree is shown in Supplementary Fig. 1. **b**, Crystal structure of the FRPL homo-dimer from *P. borbori* at 1.8 Å with head domains indicated (PDB ID 8AG8) c, Rotated overlay with FRP (PDB ID 4JDX from *Synechocystis* sp. PCC 6803)[25]. RMSD, root-mean-square deviation.

Tab. 1| Crystallographic data collection and refinement statistics

	FbFRPL
Data collection	
Space group	$P4_3\ 2_1\ 2$
Cell dimensions	
a, *b*, *c* (Å)	53.46, 53.46, 92.67
α, β, γ (°)	90, 90, 90
Resolution (Å)	46.334–1.8 (1.864–1.8)
R_{merge}	0.05548 (0.5827)
I / σI	32.02 (2.99)
Completeness (%)	98.21 (97.13)
Redundancy	22.2 (12.9)
$CC_{1/2}$	1 (0.96)
Refinement	
Resolution (Å)	35–1.8 (1.864–1.8)
No. reflections	12,829 (1,218)
R_{work} / R_{free}	0.2271 (0.2755) / 0.2322 (0.3964)
No. atoms	
Protein	877
Ligand / ion	13
Water	73
***B*-factors**	
Protein	37.94
Ligand / ion	46.87
Water	43.55
Ramachandran (%)	
Favoured	100
Allowed	0
Outliers	0
Root-mean-square deviations	
Bond lengths (Å)	0.02
Bond angles (°)	1.45

Values in parentheses are for highest-resolution shell.

FRPL evolved the ability to interact with OCP by chance

The shared fold of FRPL and FRP suggests FRPLs may be able to interact productively with OCP, meaning that they may have needed no additional modifications after being transferred into cyanobacteria to immediately function in their photoprotection system. To test this, we purified several FRPLs from extant species, and examined their effect on extant OCP1's photo-recovery. We chose FRPLs from four organisms that span the diversity of FRPL-containing bacterial groups on our phylogenetic tree: *P. borbori*, *Methylocaldum* sp. (another gamma-proteobacterium), *Chlorobi* sp. (an FCB group species) and a delta-proteobacterium of the *Desulfobacteraceae* family, which represents one of the closest extant sequences to the HGT event into cyanobacteria on our tree (Fig. 2a). FRPL from *P. borbori*, *Methylocaldum* sp. and *Chlorobi* sp. had virtually no effect on OCP1's photo-recovery. However, the *Desulfobacteraceae* FRPL showed the typical acceleration of OCP1's recovery from photoconversion by about 93% (when incubated in an equimolar ratio of OCP1 to FRPL), compared to OCP1 alone (Fig. 3a,b, Extended Data Fig. 7h–k). This indicates that the ability to regulate OCP1 already existed at the moment of the HGT event that first transferred FRP into cyanobacteria.

To further test this theory, we additionally resurrected two ancestral proteins: FRPLpreHGT that is the latest FRPL we can reconstruct before the HGT event and FRPpostHGT that represents the LCA of all FRP in cyanobacteria after the HGT (Fig. 3c). Both ancestral proteins also show the typical accelerating FRP effect on OCP1's photo-recovery, performing almost as well as extant FRP (Fig. 3d,e, Extended Data Fig. 8a–d). This inference is further robust to alternative ancestral FRP and ancestral FRPL proteins with slightly different sequences that, on the basis of an initial FRP(L) phylogeny we had inferred earlier with fewer sequences in total (Extended Data Fig. 8e–j).

Taken together, our results show that most FRPLs cannot function as allosteric regulators of OCP1, but that a small subgroup of them fortuitously acquired this ability. Because this happened in a genome that contained no OCP, this ability is entirely accidental and cannot have been the result of direct natural selection. In principle, this would have allowed the protein to function in the totally unrelated photoprotection system of cyanobacteria the moment it was first transferred into their genomes.

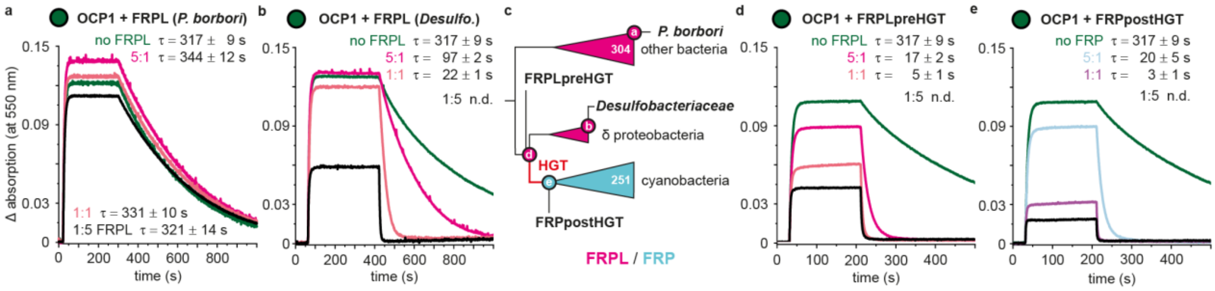


Fig. 3| Some FRPLs could fortuitously accelerate OCP's recovery from photoconversion before they were transferred into cyanobacteria. a,b, Recovery from photoconversion of extant OCP1 from *Synechocystis* sp. PCC 6803 (SYNY3) with extant FRPL of *P. borbori* (**a**) or a *Desulfobacteriaceae* (*Desulfo.*) species (**b**) at different molar ratios as indicated at 20 °C with respective mean recovery time constants (τ) and SD of three independent replicates. Representative data sets are shown for clarity. n.d., not determinable. **c**, Schematic FRP(L) phylogeny with reconstructed ancestral proteins, and extant FRPLs tested. The complete tree is shown in Supplementary Fig. 1. **d,e**, Recovery from photoconversion of extant SYNY3 OCP1 with ancestral FRPL (FRPLpreHGT) that existed before (**d**), and ancestral FRP (FRPpostHGT) that existed after the HGT (**e**) at different molar ratios as indicated at 20 °C with respective mean recovery time constants (τ) and SD of three independent replicates. Representative data sets are shown for clarity.

The OCP–FRP interface predates the allosteric accelerating effect

Since some FRPLs seem primed for the interaction with OCP even before they came into cyanobacteria, we reasoned that the interface for their interaction may also already be present in AncOCPall, even if the allosteric connection to accelerate the photo-recovery had not yet fully evolved. Analytical size-exclusion chromatography (SEC) of photoactivated, red forms of AncOCPall (AncOCPall^R) incubated with extant FRP showed increased size relative to AncOCPall^R alone (Fig. 4a), indicating that FRP already binds to AncOCPall^R.

We asked whether we could trigger the allosteric response by adding FRP in excess to the OCP^R to OCP^O recovery reaction, and repeated our initial experiments (Fig. 1d), but this time using a much larger molar ratio of FRP relative to OCP. To our surprise, instead of an acceleration, the recovery time drastically increased from 166 ± 10 to 288 ± 10 s and 609 ± 5 s, using an equimolar amount (of OCP to FRP) and a fivefold molar excess of FRP, respectively (Fig. 4b). This deceleration also appeared in AncOCP1&2, and if adding any of the ancestral FRPs or ancestral FRPLs (Fig. 4c, Extended Data Fig. 4u–x). To rule out that this slowing down is only caused by steric effects or molecular crowding, we repeated the experiments with PbFRPL (which has virtually no effect on OCP1's recovery time, even if added in molar excess: Fig. 3a), and likewise found virtually no effect on AncOCPall's recovery (Extended Data Fig. 4y).

Binding FRP alone is thus not sufficient for the accelerating allosteric effect to happen. Instead, it impedes photo-recovery of AncOCPall at high molar excess of FRP. Repetitive weak binding or an FRP that does not dissociate on the right timescale could interrupt or delay the recovery process of AncOCPall. Further, structural features on the OCP side such as the flexible linker loop between the N- and C-terminal domain or the short N-terminal extension may need to be further fine-tuned for the complex and highly efficient allosteric response of extant OCP1 to take place[16,26].

Our experiments show that the LCA of all OCPs already had a latent ability to interact with FRP, although this interaction was not yet capable of accelerating recovery. This implies that at least this interaction potential between OCP and FRP evolved purely by chance, even before these proteins first encountered each other in an ancestral cyanobacterium.

To understand the structural basis of this latent affinity, we inferred an AlphaFold2[28] model of the OCP1–FRP complex. It confidently predicted an interaction between the CTD of OCP1 and FRP (Fig. 4d, Extended Data Fig. 9a,f) that is consistent with previous small-angle X-ray scattering data[27]. The interaction exploits the same hydrophobic surface as OCP1 uses to dimerize in its red state on the phycobilisome[29]. FRP has been theorized to favor detachment of OCP1^R from the phycobilisome by down-shifting the association constant of binding and accelerating recovery by competing with this dimer interface in OCP1[27]. The residues and charges shown to be important for this dimer interface are also present in our ancestral OCPs (Extended Data Fig. 3a), potentially explaining why FRP can already interact with AncOCPall. We tested this hypothesis in two ways: first, we inferred an AlphaFold2 model of the CTD of AncOCPall, and compared its surfaces to OCP1's CTD. AncOCPall possesses the same hydrophobic surface as OCP1 with virtually all interface sites or charges identical between the two proteins. AlphaFold2 additionally predicts an interaction between this surface in AncOCPall and FRP (Fig. 4e, Extended Data Fig. 9b–e,g). Second, this model further indicates that dimerization in the red state should be an ancestral feature of all OCPs.

To test this, we used Native PAGE to understand whether our ancestral OCPs also dimerize in their activated, red form. Consistent with our prediction, activation leads to the formation of complexes consistent in size with homo-dimers in AncOCPall^R and AncOCP1^R. We did not detect red dimers in AncOCP1&2, probably due to its extremely rapid recovery time that technically impedes sustaining the red form in the gel (Fig. 4f).

Together, this indicates that the binding surface exploited by FRP is an ancient dimer interface of the red form of OCP that was already present in the LCA of all OCPs, even before FRP was recruited into the cyanobacterial system.

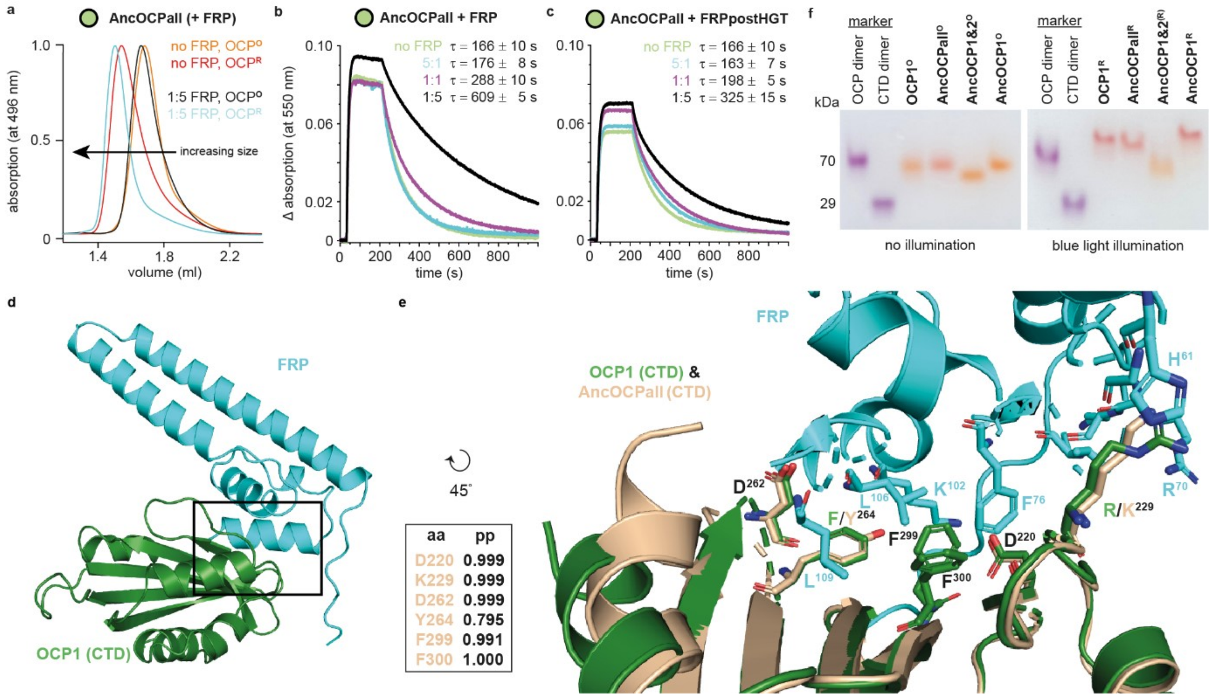


Fig. 4| Ancestral OCPs could interact with FRP through a conserved dimer interface before FRP was acquired. **a**, Analytical SEC of AncOCPall and AncOCPall–FRP complexes with (OCP^R) or without constant blue-light illumination (OCP^O) during chromatography. **b**,**c**, Recovery from photoconversion of AncOCPall with different molar ratios of extant FRP from *Synechocystis* sp. PCC 6803 (SYNY3) (**b**) or FRPpostHGT (**c**) at 20 °C with respective mean recovery time constants (τ) and SD of three independent replicates. Representative data sets are shown for clarity. Data for 'no FRP' and '5:1 FRP' in **b** are taken from Fig. 1d for comparison. **d**, AlphaFold2 model of the interaction between FRP (cyan) and the CTD of SYNY3 OCP1 (green). **e**, Rotated zoom (of black framed area in **d**) into the binding interface, with AncOCPall (in wheat) overlaid onto OCP1. Amino acids involved in binding are labelled. Sites conserved in both OCPs are in black. Nitrogen in blue and oxygen in red. Residue numbers follow SYNY3 OCP1. The insert shows the pp for indicated amino acids in the binding interface of the reconstructed AncOCPall protein. **f**, Native PAGE of ancestral OCP^O without illumination (left), and OCP^R during constant blue light illumination (right) show their oligomeric states. Comparison with OCP1[29,58] indicates conserved dimerization interfaces that differ between OCP^O and OCP^R. An OCP mutant (70 kDa) and the CTD of OCP1 (29 kDa) that both form illumination-independent dimers were used as molecular markers. Experiments were repeated three times with similar results.

OCP and FRPL drifted in and out of their ability to interact

OCPx paralogues are not affected by FRP any more[16,30]. To identify the underlying structural changes between AncOCPall and OCPx, we repeated the interaction predictions with the CTD of an extant OCPx from *Gloeobacter kilaueensis* JS1. AlphaFold2 did not predict the interaction interface between FRP and this OCPx unless we changed a conserved serine in the potential interface back to the ancestral tyrosine of AncOCPall (Extended Data Fig. 9h,i). This suggests that OCP proteins drifted in and out of the structural state that enables interaction with FRP.

To understand the structural causes of why only some FRPLs accelerate OCP1's recovery from photoconversion, we finally compared the sequences of different FRPLs. In our AlphaFold2 model, phenylalanine 76, lysine 102 and leucine 106 in FRP of *Synechocystis* sp. PCC 6803 are in contact with OCP1. Most FRPLs do not have all three states together, but occasionally have one or two of these states. *P. borbori* FRPL for instance has the phenylalanine, but features a tyrosine at position 102 and a serine at position 106 (Extended Data Fig. 8a). Other FRPLs have the lysine, but lack the phenylalanine or the leucine. This shows that the important states for the interaction with OCP1 individually come and go across the FRPL phylogeny. All three states only appeared together in FRPLs along the linage towards delta-proteobacteria and cyanobacteria. It is remarkable that the HGT into cyanobacteria happened exactly in this narrow window of full compatibility.

Discussion

Here, we have reconstructed the evolution of an allosteric interaction in the cyanobacterial photoprotection system. Together with previous work on the initial evolution of OCP[13,16], the picture that emerges is a remarkable example of evolutionary tinkering[31]: OCPs were most likely created by a gene fusion event that required nothing but a flexible linker to create a photo-switchable protein out of two non-switchable components[16]. Horizontal acquisition of FRP then introduced a new component that could allosterically accelerate ground state recovery in OCP1 without any further modification. Creating the fully functional OCP1–FRP system then only required substitutions in OCP that converted an initially unproductive interaction with the CTD into one that results in an acceleration of photo-recovery (Fig. 5). Because we cannot time the acquisition of FRP precisely relative to our OCP ancestors, we do not know whether these substitutions occurred before or after FRP was acquired. If they had happened before, the regulatory interaction between OCP1 and FRP would have been completely functional the moment FRP was horizontally acquired.

Another known function of FRP is the facilitation of OCP1 detachment from the phycobilisome by shifting the OCP^R–phycobilisome binding equilibrium constant[15]. Although this aspect was not surveyed in our study, we imagine that competitive FRP binding to an ancestral OCPR dimer could also facilitate the detachment from the phycobilisome or at least impede binding to it, in effect generating a potential ancestral mode of regulation that could have also been functional the moment FRP first appeared in cyanobacteria.

One question that remains is why was FRP recruited into the cyanobacterial photoprotection system at all? OCPs that existed before FRP was recruited could recover quickly on their own. Why complicate this functional system? We are aware of two postulated adaptive benefits: first, the OCP1–FRP interaction may offer more sophisticated control of energy use in fast-changing light regimes in the cyanobacterial cell[13]. OCP-mediated photoprotection systems without FRP can only be regulated on the level of messenger RNA transcripts, which act only slowly on a return from stressful to normal light conditions, whereas control by FRP allows potentially faster posttranslational regulation[32]. Second, it may afford superior photoprotection in high light conditions: OCP2 and OCPx paralogues recover so fast that they struggle to stably accumulate the red form at room temperature[13]. OCP1's more stable red state may then be useful when large amounts of active OCP^R are needed, but this high stability may come at the expense of being unable to recover alone. In this scenario, the recruitment of FRP would have enabled the evolution of an ultimately more efficient photoprotection mechanism.

However, the interaction could also be an example of non-adaptive complexity that simply became difficult to lose[33]: the acquisition of FRP may have enabled OCP1 to 'forget' how to recover efficiently on its own. Once it had lost this ability, FRP would have become essential for full OCP1 function.

The specific compatibility of the FRPL from the *Desulfobacteraceae* species with cyanobacterial OCPs is entirely accidental, because this protein evolved in a genome that contains no OCP. This proves that highly complementary protein surfaces can evolve completely by chance, and that such initially accidental interactions can become incorporated into the biology of organisms. Our work thus raises the possibility that some or even many protein–protein interactions are initially created without the action of direct natural selection. Organisms may in fact be bombarded with virtually fully formed interactions that are created when horizontal transfer, changes in cellular localization or spatiotemporal expression patterns bring together proteins with fortuitously compatible surfaces. From this pool, natural selection would then purge those that are harmful, fix those that are useful and ignore those that are harmless.

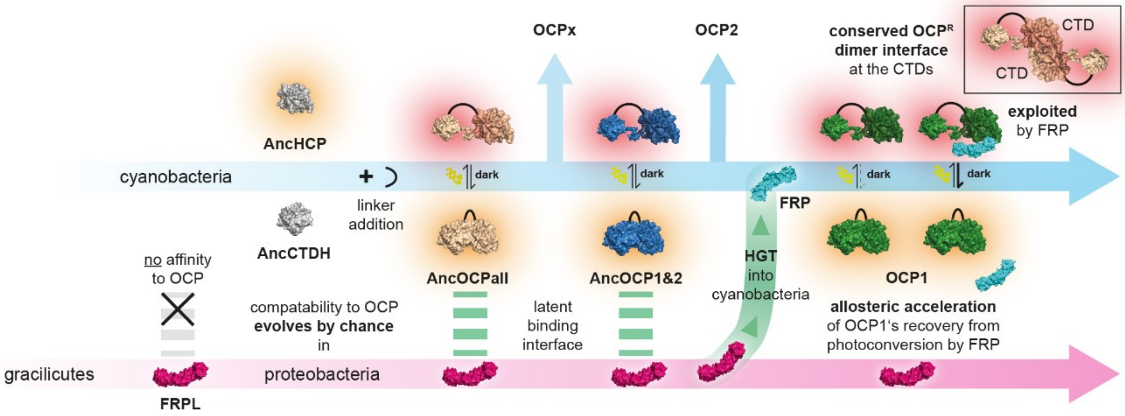


Fig. 5| Evolutionary origin of the allosteric regulator FRP in the cyanobacterial OCP-mediated photoprotection system. The first photo-switchable OCP that undergoes conformational change from a closed orange to an open red state on high light irradiation was formed in a fusion event of an ancestral HCP (AncHCP) and an ancestral CTD-like homologue (AncCTDH) via a linker addition[16]. An FRP-like protein (FRPL) was horizontally transferred (HGT) into the unrelated cyanobacterial system after a latent binding interface for ancestral OCPs had already evolved by chance. FRP now exploits the conserved CTD dimerization interface of OCP^R to strongly accelerate OCP1's recovery from photoconversion. OCP structure used here for illustration only is PDB ID 3MG1[58].

Supplementary Information

Supplementary Discussion

The fortuitous ability of FRPL to act on OCP is agnostic about an HGT event

We postulate that cyanobacteria acquired FRP via a horizontal gene transfer (HGT) from delta-proteobacteria early in their history after OCP first formed in an ancestral cyanobacterium (Fig. 2a). This is the most parsimonious explanation given the available sequence data our phylogenetic tree is based on, but also implies a large number of gene losses in different kinds of non-cyanobacterial groups (Supplementary Fig. 3a). A root between FRPs and FRPLs would place the evolution of FRPLs near the last common ancestor (LCA) of all bacteria[23], which would imply even more gene losses in virtually all major bacterial groups (Supplementary Fig. 3b). However, even in this scenario, the intrinsic ability of FRPLs to act on OCP would have evolved in a non-cyanobacterium by chance without direct selective pressure. In fact, FRPLs would have randomly drifted in and out of the sequence space that enables the interaction with OCP during its evolution, and happened to be capable of the interaction when OCP first evolved in cyanobacteria.

We have also considered the possibility that the root may lie within the FRP clade, implying HGT to proteobacteria and acidobacteria. However, we consider this scenario very implausible: it would require a transfer from a relatively modern cyanobacterium with FRP into at least the LCA of all proteobacteria and acidobacteria. This is not only temporally implausible, but further incongruent with the topology of our gene tree: when rooted inside FRPs, our phylogeny does not place proteobacteria and acidobacteria sister to all FRPs. This means our gene tree would require additional HGT between different kinds of proteobacteria to explain the distribution of FRPL (Supplementary Fig. 3c). In addition, this scenario implies that FRPLs in delta-proteobacteria fortuitously retained their ability to interact with OCP since around the time of the LCA of all living cyanobacteria.

Another possibility is that our gene tree is simply incorrect, perhaps owing to the short length of the FRP and FRPL proteins. For example, if the true tree in the FRPL clade actually follows the species phylogeny of acidobacteria and proteobacteria, we could root the tree inside the FRP clade and explain the tree with a single horizontal transfer (Supplementary Fig. 3d). FRP's function would then be ancestral. But this would also imply that the ability to bind OCP was lost many times independently in

FRPLs, and was regained in only a small set of delta-proteobacteria. We also consider this very implausible. Further, a reconciliation of 100 bootstrap trees with a species phylogeny using amalgamated likelihood estimation (ALE)[45] found no root between FRPs and FRPLs, but 72 tree topologies featuring a root within the FRPLs.

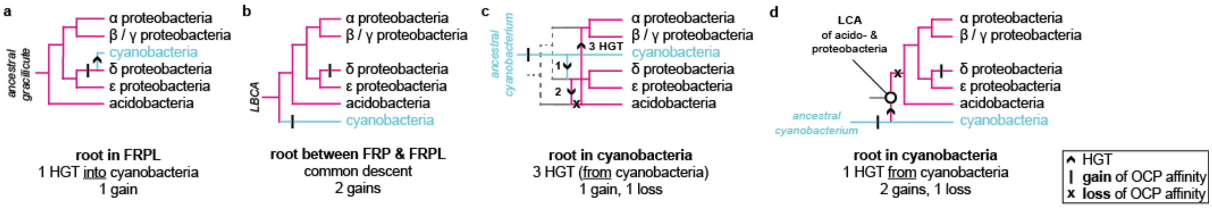
Taken together, the OCP-matching interface in FRPLs evolved without direct selective pressure mainly by chance even if we remain agnostic about the horizontal transfer event, that we think is still the most likely scenario here.

Supplementary Data 1| See online version of the manuscript for the data.

Supplementary Tab. 1| See online version of the manuscript for the table.

Supplementary Fig. 1| See online version of the manuscript for the figure.

Supplementary Fig. 2| See online version of the manuscript for the figure.



Supplementary Fig. 3| Scenarios of FRP and FRPL evolution. a, Simplified scheme of the evolution of FRP and FRPL proteins, according to the most parsimonious scenario shown in Fig. 2a. **b-d**, Simplified schemes of the alternative, but less parsimonious scenarios. *LBCA*, last bacterial common ancestor. HGT, horizontal gene transfer.

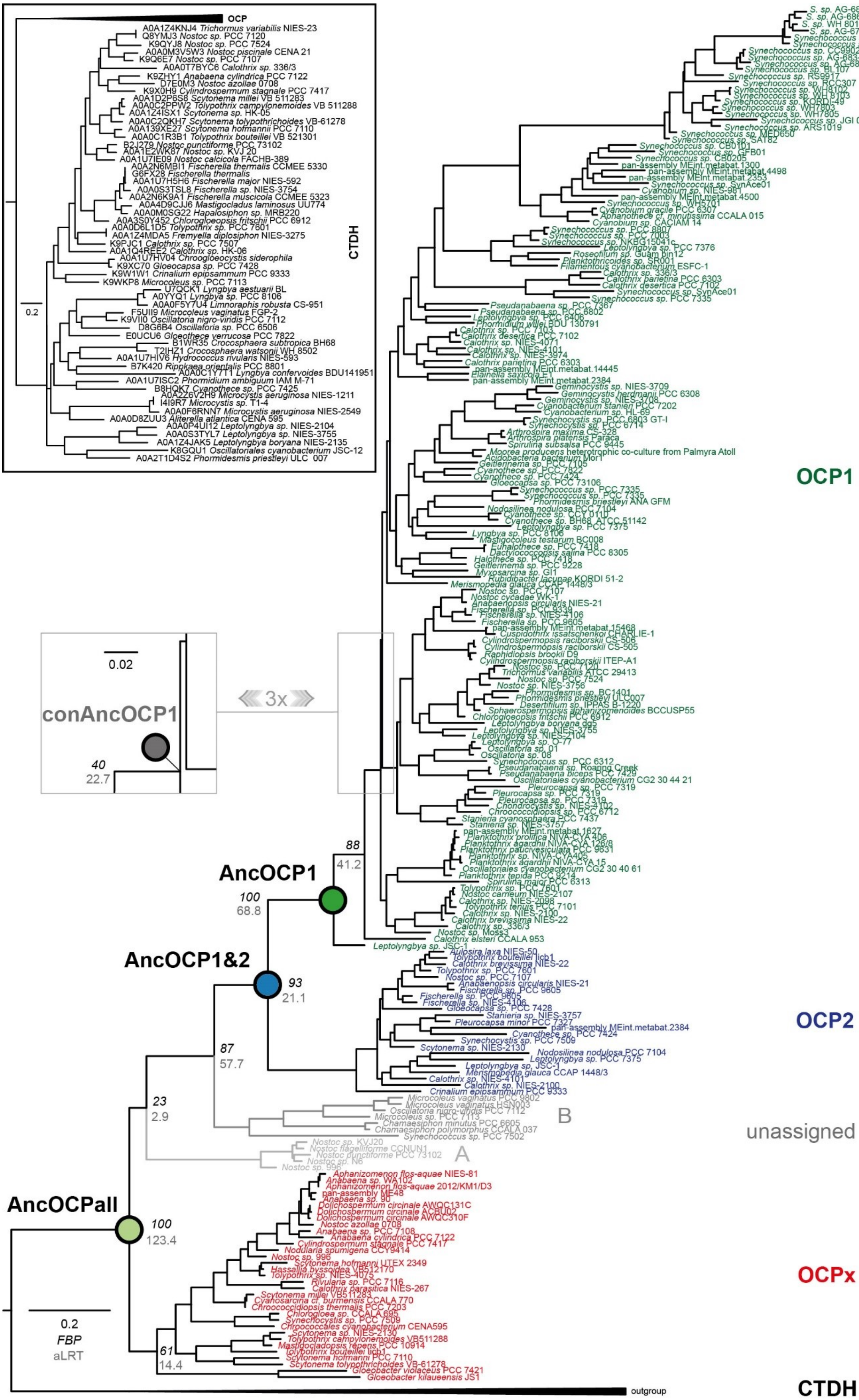
Extended Data Figures

Extended Data Fig. 1| See following page for figure. Caption is below.

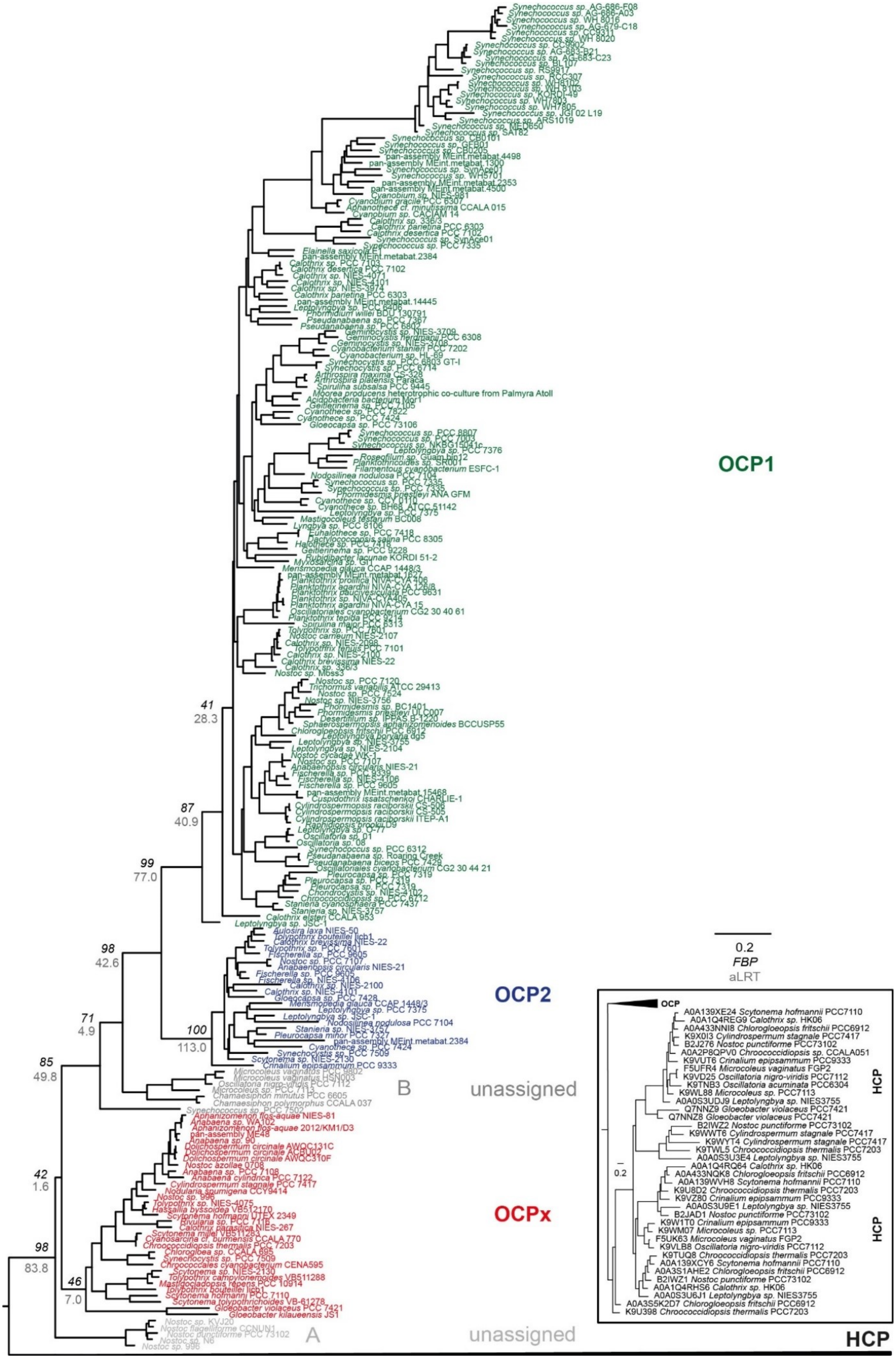
Extended Data Fig. 2| See second following page for figure. Caption is below.

Extended Data Fig. 1| Complete phylogeny of OCP proteins. ML phylogeny of OCP proteins with reconstructed ancestral proteins (Anc) at labelled nodes, and cyanobacterial C-terminal domain-like proteins (CTDH) as the outgroup (insert with black outlines). OCP paralogs and ancestors are color-coded as in Fig. 1b. We additionally tested a more conservative sequence for the last common ancestor of OCP1 (conAncOCP1, in grey) (Extended Data Fig. 3a+e, 4d,h,l,p,t) as well as alternative 'altAll' ancestors for every node on this tree (Extended Data Figs. 3a, 5a-l). Italic numbers are Felsenstein Bootstrap Probabilities (FBP) of 100 replicates. Grey numbers are approximate likelihood-ratio test values (aLRT). Branch-lengths represent average substitutions per site. Insert with grey outlines is a threefold zoom-in to properly display the branch topology in that area. Underlying multiple sequence alignment in Supplementary Data 1.

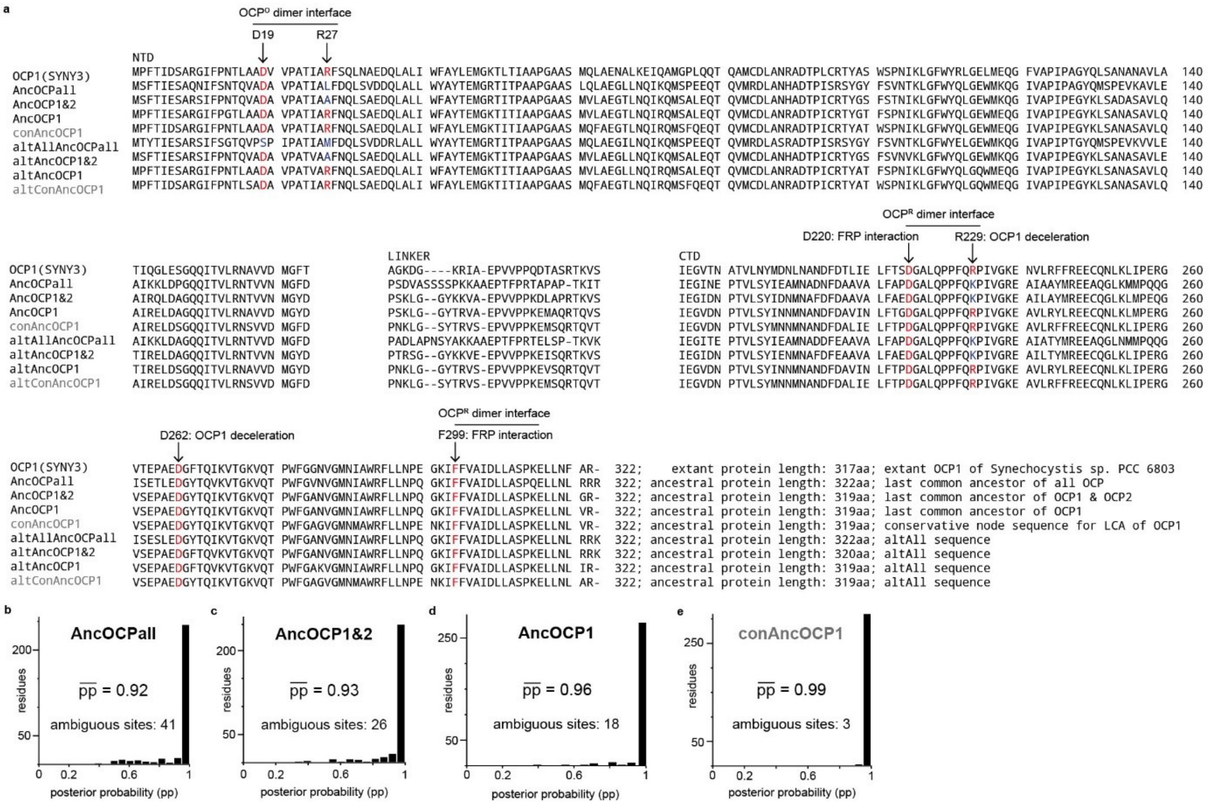
Extended Data Fig. 2| Alternatively rooted phylogeny of OCP proteins. ML phylogeny of OCP proteins like in Extended Data Fig. 1, but with cyanobacterial helical carotenoid proteins (HCP, insert) as the outgroup. Underlying multiple sequence alignment in Supplementary Data 1. No ancestors were reconstructed here.


OCP
A0A1Z4KNJ4 Trichormus variabilis NIES-23
Q8YMJ3 Nostoc sp. PCC 7120
K9QYJ8 Nostoc sp. PCC 7524
A0A0M3V5W3 Nostoc piscinale CENA 21
K9Q6E7 Nostoc sp. PCC 7107
A0A0T7BYC6 Calothrix sp. 336/3
K9ZHY1 Anabaena cylindrica PCC 7122
D7E0M3 Nostoc azollae 0708
K9X0H9 Cylindrospermum stagnale PCC 7417
A0A1D2P6S8 Scytonema millei VB 511283
A0A0C2PPW2 Tolypothrix campylonemoides VB 511288
A0A1Z4ISX1 Scytonema sp. HK-05
A0A0C2QKH7 Scytonema tolypothrichoides VB-61278
A0A139XE27 Scytonema hofmannii PCC 7110
A0A0C1R3B1 Tolypothrix bouteillei VB 521301
B2J279 Nostoc punctiforme PCC 73102
A0A1E2WK87 Nostoc sp. KVJ 20
A0A1U7IE09 Nostoc calcicola FACHB-389
A0A2N6MBI1 Fischerella thermalis CCMEE 5330
G6FX28 Fischerella thermalis
A0A1U7H5H6 Fischerella major NIES-592
A0A0S3TSL8 Fischerella sp. NIES-3754
A0A2N6K9A1 Fischerella muscicola CCMEE 5323
A0A4D9CJJ6 Mastigocladus laminosus UU774
A0A0M0SG22 Hapalosiphon sp. MRB220
A0A3S0Y452 Chlorogloeopsis fritschii PCC 6912
A0A0D6L1D5 Tolypothrix sp. PCC 7601
A0A1Z4MDA5 Fremyella diplosiphon NIES-3275
K9PJC1 Calothrix sp. PCC 7507
A0A1Q4REE2 Calothrix sp. HK-06
A0A1U7HV04 Chroogloeocystis siderophila
K9XC70 Gloeocapsa sp. PCC 7428
K9W1W1 Crinalium epipsammum PCC 9333
K9WKP8 Microcoleus sp. PCC 7113
U7QCK1 Lyngbya aestuarii BL
A0YYQ1 Lyngbya sp. PCC 8106
A0A0F5Y7U4 Limnoraphis robusta CS-951
F5UII9 Microcoleus vaginatus FGP-2
K9VII0 Oscillatoria nigro-viridis PCC 7112
D8G6B4 Oscillatoria sp. PCC 6506
E0UCU6 Gloeothece verrucosa PCC 7822
B1WR35 Crocosphaera subtropica BH68
T2IHZ1 Crocosphaera watsonii WH 8502
A0A1U7HIV6 Hydrococcus rivularis NIES-593
B7K420 Rippkaea orientalis PCC 8801
A0A0C1Y7T1 Lyngbya confervoides BDU141951
A0A1U7ISC2 Phormidium ambiguum IAM M-71
B8HQK7 Cyanothece sp. PCC 7425
A0A2Z6V2H9 Microcystis aeruginosa NIES-1211
I4I9R7 Microcystis sp. T1-4
A0A0F6RNN7 Microcystis aeruginosa NIES-2549
A0A0D8ZUU3 Aliterella atlantica CENA 595
A0A0P4UI12 Leptolyngbya sp. NIES-2104
A0A0S3TYL7 Leptolyngbya sp. NIES-3755
A0A1Z4JAK5 Leptolyngbya boryana NIES-2135
K8GQU1 Oscillatoriales cyanobacterium JSC-12
A0A2T1D4S2 Phormidesmis priestleyi ULC 007
CTDH
0.2
S. sp. AG-686-F08
S. sp. AG-686-A03
S. sp. WH 8016
S. sp. AG-679-C18
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Synechococcus sp. CC9902
Synechococcus sp. AG-683-B21
Synechococcus sp. AG-683-C23
Synechococcus sp. BL107
Synechococcus sp. RS9917
Synechococcus sp. RCC307
Synechococcus sp. WH8102
Synechococcus sp. WH 8103
Synechococcus sp. KORDI-49
Synechococcus sp. WH7803
Synechococcus sp. WH7805
Synechococcus sp. JGI 02 L19
Synechococcus sp. ARS1019
Synechococcus sp. MED650
Synechococcus sp. SAT82
Synechococcus sp. CB0101
Synechococcus sp. GFB01
Synechococcus sp. CB0205
pan-assembly MEint.metabat.1300
pan-assembly MEint.metabat.4498
pan-assembly MEint.metabat.2353
Synechococcus sp. SynAce01
Cyanobium sp. NIES-981
pan-assembly MEint.metabat.4500
Synechococcus sp. WH5701
Cyanobium gracile PCC 6307
Aphanothece cf. minutissima CCALA 015
Cyanobium sp. CACIAM 14
Synechococcus sp. PCC 8807
Synechococcus sp. PCC 7003
Synechococcus sp. NKBG15041c
Leptolyngbya sp. PCC 7376
Roseofilum sp. Guam bin12
Planktothricoides sp. SR001
Filamentous cyanobacterium ESFC-1
Calothrix sp. 336/3
Calothrix parietina PCC 6303
Calothrix desertica PCC 7102
Synechococcus sp. SynAce01
Synechococcus sp. PCC 7335
Pseudanabaena sp. PCC 7367
Pseudanabaena sp. PCC 6802
Leptolyngbya sp. PCC 6406
Phormidium willei BDU 130791
Calothrix sp. PCC 7103
Calothrix desertica PCC 7102
Calothrix sp. NIES-4071
Calothrix sp. NIES-4101
Calothrix sp. NIES-3974
Calothrix parietina PCC 6303
pan-assembly MEint.metabat.14445
Elainella saxicola E1
pan-assembly MEint.metabat.2384
Geminocystis sp. NIES-3709
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES-3708
Cyanobacterium stanieri PCC 7202
Cyanobacterium sp. HL-69
Synechocystis sp. PCC 6803 GT-I
Synechocystis sp. PCC 6714
Arthrospira maxima CS-328
Arthrospira platensis Paraca
Spirulina subsalsa PCC 9445
Moorea producens heterotrophic co-culture from Palmyra Atoll
Acidobacteria bacterium Mor1
Geitlerinema sp. PCC 7105
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Gloeocapsa sp. PCC 73106
Synechococcus sp. PCC 7335
Synechococcus sp. PCC 7335
Phormidesmis priestleyi ANA GFM
Nodosilinea nodulosa PCC 7104
Cyanothece sp. CCY 0110
Cyanothece sp. BH68, ATCC 51142
Leptolyngbya sp. PCC 7375
Lyngbya sp. PCC 8106
Mastigocoleus testarum BC008
Euhalothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Geitlerinema sp. PCC 9228
Myxosarcina sp. GI1
Rubidibacter lacunae KORDI 51-2
Merismopedia glauca CCAP 1448/3
Nostoc sp. PCC 7107
Nostoc cycadae WK-1
Anabaenopsis circularis NIES-21
Fischerella sp. PCC 9339
Fischerella sp. NIES-4106
Fischerella sp. PCC 9605
pan-assembly MEint.metabat.15468
Cuspidothrix issatschenkoi CHARLIE-1
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-505
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii ITEP-A1
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. NIES-3756
Phormidesmis sp. BC1401
Phormidesmis priestleyi ULC007
Desertifilum sp. IPPAS B-1220
Sphaerospermopsis aphanizomenoides BCCUSP55
Chlorogloeopsis fritschii PCC 6912
Leptolyngbya boryana dg5
Leptolyngbya sp. NIES-3755
Leptolyngbya sp. NIES-2104
Leptolyngbya sp. O-77
Oscillatoria sp. 01
Oscillatoria sp. 08
Synechococcus sp. PCC 6312
Pseudanabaena sp. Roaring Creek
Pseudanabaena biceps PCC 7429
Oscillatoriales cyanobacterium CG2 30 44 21
Pleurocapsa sp. PCC 7319
Pleurocapsa sp. PCC 7319
Pleurocapsa sp. PCC 7319
Chondrocystis sp. NIES-4102
Chroococcidiopsis sp. PCC 6712
Stanieria cyanosphaera PCC 7437
Stanieria sp. NIES-3757
pan-assembly MEint.metabat.1627
Planktothrix prolifica NIVA-CYA 406
Planktothrix agardhii NIVA-CYA 126/8
Planktothrix paucivesiculata PCC 9631
Planktothrix sp. NIVA-CYA405
Planktothrix agardhii NIVA-CYA 15
Oscillatoriales cyanobacterium CG2 30 40 61
Planktothrix tepida PCC 9214
Spirulina major PCC 6313
Tolypothrix sp. PCC 7601
Nostoc carneum NIES-2107
Calothrix sp. NIES-2098
Tolypothrix tenuis PCC 7101
Calothrix sp. NIES-2100
Calothrix brevissima NIES-22
Calothrix sp. 336/3
Nostoc sp. Moss3
Calothrix elsteri CCALA 953
Leptolyngbya sp. JSC-1
OCP1
0.02
conAncOCP1
40
22.7
3x
88
41.2
AncOCP1
100
68.8
AncOCP1&2
93
21.1
Aulosira laxa NIES-50
Tolypothrix bouteillei licb1
Calothrix brevissima NIES-22
Tolypothrix sp. PCC 7601
Nostoc sp. PCC 7107
Anabaenopsis circularis NIES-21
Fischerella sp. PCC 9605
Fischerella sp. PCC 9605
Fischerella sp. NIES-4106
Gloeocapsa sp. PCC 7428
Stanieria sp. NIES-3757
Pleurocapsa minor PCC 7327
pan-assembly MEint.metabat.2384
Cyanothece sp. PCC 7424
Synechocystis sp. PCC 7509
Scytonema sp. NIES-2130
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp. PCC 7375
Leptolyngbya sp. JSC-1
Merismopedia glauca CCAP 1448/3
Calothrix sp. NIES-4101
Calothrix sp. NIES-2100
Crinalium epipsammum PCC 9333
OCP2
87
57.7
23
2.9
Microcoleus vaginatus PCC 9802
Microcoleus vaginatus HSN003
Oscillatoria nigro-viridis PCC 7112
Microcoleus sp. PCC 7113
Chamaesiphon minutus PCC 6605
Chamaesiphon polymorphus CCALA 037
Synechococcus sp. PCC 7502
B
unassigned
Nostoc sp. KVJ20
Nostoc flagelliforme CCNUN1
Nostoc punctiforme PCC 73102
Nostoc sp. N6
Nostoc sp. 996
A
AncOCPall
100
123.4
Aphanizomenon flos-aquae NIES-81
Anabaena sp. WA102
Aphanizomenon flos-aquae 2012/KM1/D3
pan-assembly ME48
Anabaena sp. 90
Dolichospermum circinale AWQC131C
Dolichospermum circinale ACBU02
Dolichospermum circinale AWQC310F
Nostoc azollae 0708
Anabaena sp. PCC 7108
Anabaena cylindrica PCC 7122
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. 996
Scytonema hofmanni UTEX 2349
Hassallia byssoidea VB512170
Tolypothrix sp. NIES-4075
Rivularia sp. PCC 7116
Calothrix parasitica NIES-267
Scytonema millei VB511283
Cyanosarcina cf. burmensis CCALA 770
Chroococcidiopsis thermalis PCC 7203
Chlorogloea sp. CCALA 695
Synechocystis sp. PCC 7509
Chroococcales cyanobacterium CENA595
Scytonema sp. NIES-2130
Tolypothrix campylonemoides VB511288
Mastigocladopsis repens PCC 10914
Tolypothrix bouteillei licb1
Scytonema hofmanni PCC 7110
Scytonema tolypothrichoides VB-61278
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
OCPx
0.2
FBP
aLRT
61
14.4
outgroup
CTDH

Extended Data Fig. 1| See previous page for caption.


Synechococcus sp. AG-686-F08
Synechococcus sp. AG-686-A03
Synechococcus sp. WH 8016
Synechococcus sp. AG-679-C18
Synechococcus sp. CC9311
Synechococcus sp. WH 8020
Synechococcus sp. CC9902
Synechococcus sp. AG-683-B21
Synechococcus sp. AG-683-C23
Synechococcus sp. BL107
Synechococcus sp. RS9917
Synechococcus sp. RCC307
Synechococcus sp. WH8102
Synechococcus sp. WH 8103
Synechococcus sp. KORDI-49
Synechococcus sp. WH7803
Synechococcus sp. WH7805
Synechococcus sp. JGI 02 L19
Synechococcus sp. ARS1019
Synechococcus sp. MED650
Synechococcus sp. SAT82
Synechococcus sp. CB0101
Synechococcus sp. GFB01
Synechococcus sp. CB0205
pan-assembly MEint.metabat.4498
pan-assembly MEint.metabat.1300
Synechococcus sp. SynAce01
Synechococcus sp. WH5701
pan-assembly MEint.metabat.2353
pan-assembly MEint.metabat.4500
Cyanobium sp. NIES-981
Cyanobium gracile PCC 6307
Aphanothece cf. minutissima CCALA 015
Cyanobium sp. CACIAM 14
Calothrix sp. 336/3
Calothrix parietina PCC 6303
Calothrix desertica PCC 7102
Synechococcus sp. SynAce01
Synechococcus sp. PCC 7335
Elainella saxicola E1
pan-assembly MEint.metabat.2384
Calothrix sp. PCC 7103
Calothrix desertica PCC 7102
Calothrix sp. NIES-4071
Calothrix sp. NIES-4101
Calothrix sp. NIES-3974
Calothrix parietina PCC 6303
pan-assembly MEint.metabat.14445
Leptolyngbya sp. PCC 6406
Phormidium willei BDU 130791
Pseudanabaena sp. PCC 7367
Pseudanabaena sp. PCC 6802
Geminocystis sp. NIES-3709
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES-3708
Cyanobacterium stanieri PCC 7202
Cyanobacterium sp. HL-69
Synechocystis sp. PCC 6803 GT-I
Synechocystis sp. PCC 6714
Arthrospira maxima CS-328
Arthrospira platensis Paraca
Spirulina subsalsa PCC 9445
Moorea producens heterotrophic co-culture from Palmyra Atoll
Acidobacteria bacterium Mor1
Geitlerinema sp. PCC 7105
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Gloeocapsa sp. PCC 73106
Synechococcus sp. PCC 8807
Synechococcus sp. PCC 7003
Synechococcus sp. NKBG15041c
Leptolyngbya sp. PCC 7376
Roseofilum sp. Guam bin12
Planktothricoides sp. SR001
Filamentous cyanobacterium ESFC-1
Nodosilinea nodulosa PCC 7104
Synechococcus sp. PCC 7335
Synechococcus sp. PCC 7335
Phormidesmis priestleyi ANA GFM
Cyanothece sp. CCY 0110
Cyanothece sp. BH68, ATCC 51142
Leptolyngbya sp. PCC 7375
Mastigocoleus testarum BC008
Lyngbya sp. PCC 8106
Euhalothece sp. PCC 7418
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Geitlerinema sp. PCC 9228
Rubidibacter lacunae KORDI 51-2
Myxosarcina sp. GI1
Merismopedia glauca CCAP 1448/3
pan-assembly MEint.metabat.1627
Planktothrix prolifica NIVA-CYA 406
Planktothrix agardhii NIVA-CYA 126/8
Planktothrix paucivesiculata PCC 9631
Planktothrix sp. NIVA-CYA405
Planktothrix agardhii NIVA-CYA 15
Oscillatoriales cyanobacterium CG2 30 40 61
Planktothrix tepida PCC 9214
Spirulina major PCC 6313
Tolypothrix sp. PCC 7601
Nostoc carneum NIES-2107
Calothrix sp. NIES-2098
Tolypothrix tenuis PCC 7101
Calothrix sp. NIES-2100
Calothrix brevissima NIES-22
Calothrix sp. 336/3
Nostoc sp. Moss3
Nostoc sp. PCC 7120
Trichormus variabilis ATCC 29413
Nostoc sp. PCC 7524
Nostoc sp. NIES-3756
Phormidesmis sp. BC1401
Phormidesmis priestleyi ULC007
Desertifilum sp. IPPAS B-1220
Sphaerospermopsis aphanizomenoides BCCUSP55
Chlorogloeopsis fritschii PCC 6912
Leptolyngbya boryana dg5
Leptolyngbya sp. NIES-3755
Leptolyngbya sp. NIES-2104
Nostoc cycadae WK-1
Nostoc sp. PCC 7107
Anabaenopsis circularis NIES-21
Fischerella sp. PCC 9339
Fischerella sp. NIES-4106
Fischerella sp. PCC 9605
pan-assembly MEint.metabat.15468
Cuspidothrix issatschenkoi CHARLIE-1
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-505
Cylindrospermopsis raciborskii ITEP-A1
Raphidiopsis brookii D9
Leptolyngbya sp. O-77
Oscillatoria sp. 01
Oscillatoria sp. 08
Synechococcus sp. PCC 6312
Pseudanabaena sp. Roaring Creek
Pseudanabaena biceps PCC 7429
Oscillatoriales cyanobacterium CG2 30 44 21
Pleurocapsa sp. PCC 7319
Pleurocapsa sp. PCC 7319
Pleurocapsa sp. PCC 7319
Chondrocystis sp. NIES-4102
Chroococcidiopsis sp. PCC 6712
Stanieria cyanosphaera PCC 7437
Stanieria sp. NIES-3757
Calothrix elsteri CCALA 953
Leptolyngbya sp. JSC-1
OCP1
41
28.3
87
40.9
99
77.0
Aulosira laxa NIES-50
Tolypothrix bouteillei licb1
Calothrix brevissima NIES-22
Tolypothrix sp. PCC 7601
Fischerella sp. PCC 9605
Nostoc sp. PCC 7107
Anabaenopsis circularis NIES-21
Fischerella sp. PCC 9605
Fischerella sp. NIES-4106
Calothrix sp. NIES-2100
Calothrix sp. NIES-4101
Gloeocapsa sp. PCC 7428
Merismopedia glauca CCAP 1448/3
Leptolyngbya sp. PCC 7375
Leptolyngbya sp. JSC-1
Nodosilinea nodulosa PCC 7104
Stanieria sp. NIES-3757
Pleurocapsa minor PCC 7327
pan-assembly MEint.metabat.2384
Cyanothece sp. PCC 7424
Synechocystis sp. PCC 7509
Scytonema sp. NIES-2130
Crinalium epipsammum PCC 9333
OCP2
100
113.0
98
42.6
71
4.9
85
49.8
Microcoleus vaginatus PCC 9802
Microcoleus vaginatus HSN003
Oscillatoria nigro-viridis PCC 7112
Microcoleus sp. PCC 7113
Chamaesiphon minutus PCC 6605
Chamaesiphon polymorphus CCALA 037
B
unassigned
Synechococcus sp. PCC 7502
Aphanizomenon flos-aquae NIES-81
Anabaena sp. WA102
Aphanizomenon flos-aquae 2012/KM1/D3
pan-assembly ME48
Anabaena sp. 90
Dolichospermum circinale AWQC131C
Dolichospermum circinale ACBU02
Dolichospermum circinale AWQC310F
Nostoc azollae 0708
Anabaena sp. PCC 7108
Anabaena cylindrica PCC 7122
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc sp. 996
Tolypothrix sp. NIES-4075
Hassallia byssoidea VB512170
Scytonema hofmanni UTEX 2349
Rivularia sp. PCC 7116
Calothrix parasitica NIES-267
Scytonema millei VB511283
Cyanosarcina cf. burmensis CCALA 770
Chroococcidiopsis thermalis PCC 7203
Chlorogloea sp. CCALA 695
Synechocystis sp. PCC 7509
Chroococcales cyanobacterium CENA595
Scytonema sp. NIES-2130
Tolypothrix campylonemoides VB511288
Mastigocladopsis repens PCC 10914
Tolypothrix bouteillei licb1
Scytonema hofmanni PCC 7110
Scytonema tolypothrichoides VB-61278
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
OCPx
42
1.6
46
7.0
98
83.8
Nostoc sp. KVJ20
Nostoc flagelliforme CCNUN1
Nostoc punctiforme PCC 73102
Nostoc sp. N6
Nostoc sp. 996
A
unassigned
0.2
FBP
aLRT
OCP
A0A139XE24 Scytonema hofmannii PCC7110
A0A1Q4REG9 Calothrix sp. HK06
A0A433NNI8 Chlorogloeopsis fritschii PCC6912
K9X0I3 Cylindrospermum stagnale PCC7417
B2J276 Nostoc punctiforme PCC73102
A0A2P8QPV0 Chroococcidiopsis sp. CCALA051
K9VUT6 Crinalium epipsammum PCC9333
F5UFR4 Microcoleus vaginatus FGP2
K9VD25 Oscillatoria nigro-viridis PCC7112
K9TNB3 Oscillatoria acuminata PCC6304
K9WL88 Microcoleus sp. PCC7113
A0A0S3UDJ9 Leptolyngbya sp. NIES3755
Q7NNZ9 Gloeobacter violaceus PCC7421
Q7NNZ8 Gloeobacter violaceus PCC7421
HCP
B2IWZ2 Nostoc punctiforme PCC73102
K9WWT6 Cylindrospermum stagnale PCC7417
K9WYT4 Cylindrospermum stagnale PCC7417
K9TWL5 Chroococcidiopsis thermalis PCC7203
A0A0S3U3E4 Leptolyngbya sp. NIES3755
0.2
A0A1Q4RQ64 Calothrix sp. HK06
A0A433NQK8 Chlorogloeopsis fritschii PCC6912
A0A139WVH8 Scytonema hofmannii PCC7110
K9U8D2 Chroococcidiopsis thermalis PCC7203
K9VZ80 Crinalium epipsammum PCC9333
A0A0S3U9E1 Leptolyngbya sp. NIES3755
B2JAD1 Nostoc punctiforme PCC73102
K9W1T0 Crinalium epipsammum PCC9333
K9WM07 Microcoleus sp. PCC7113
F5UK63 Microcoleus vaginatus FGP2
K9VLB8 Oscillatoria nigro-viridis PCC7112
HCP
K9TUQ8 Chroococcidiopsis thermalis PCC7203
A0A139XCY6 Scytonema hofmannii PCC7110
A0A3S1AHE2 Chlorogloeopsis fritschii PCC6912
B2IWZ1 Nostoc punctiforme PCC73102
A0A1Q4RHS6 Calothrix sp. HK06
A0A0S3U6J1 Leptolyngbya sp. NIES3755
A0A3S5K2D7 Chlorogloeopsis fritschii PCC6912
K9U398 Chroococcidiopsis thermalis PCC7203
HCP

Extended Data Fig. 2| See second previous page for caption.

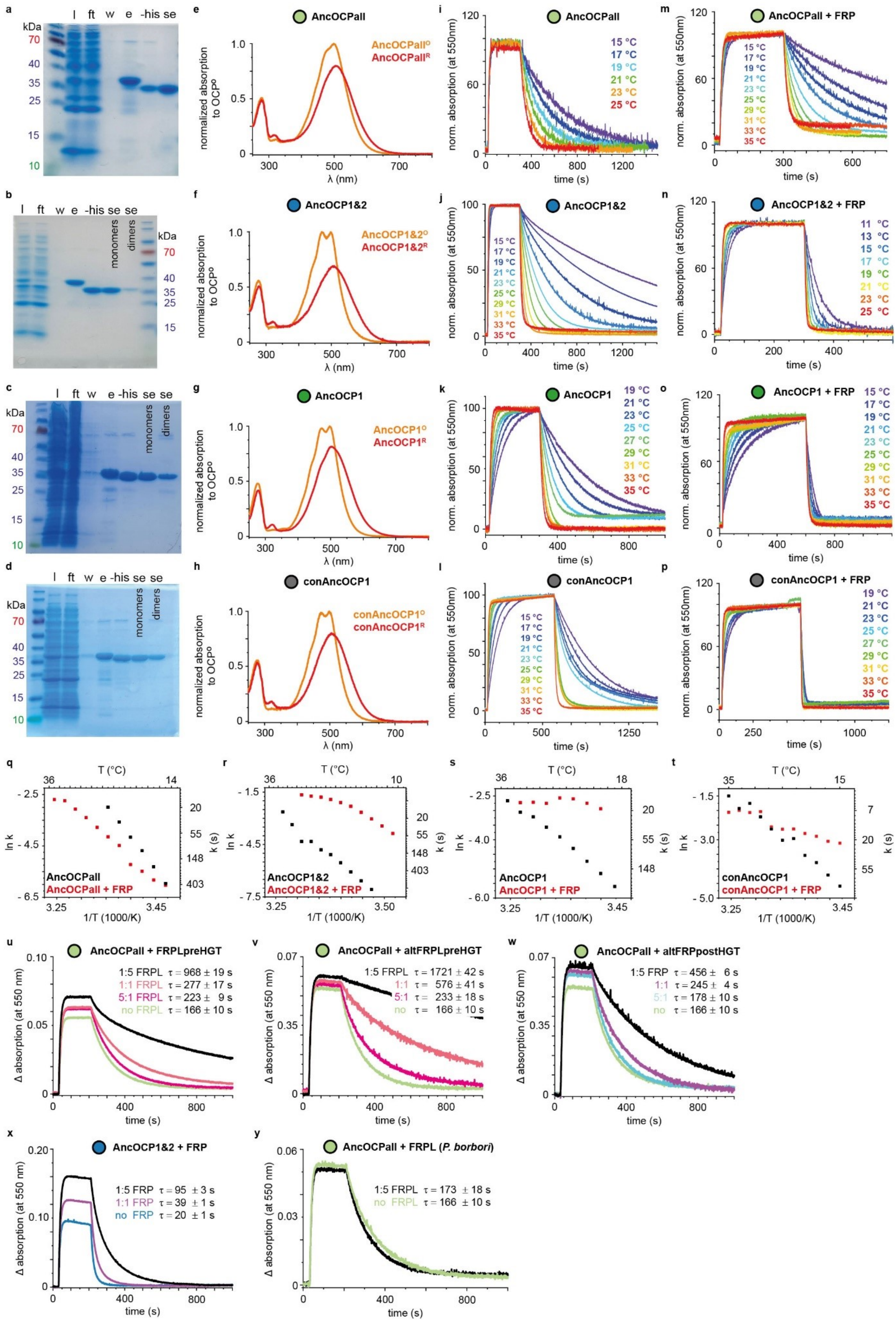


```
                  NTD
OCP1(SYNY3)       MPFTIDSARGIFPNTLAADV VPATIARFSQLNAEDQLALI WFAYLEMGKTLTIAAPGAAS MQLAENALKEIQAMGPLQQT QAMCDLANRADTPLCRTYAS WSPNIKLGFWYRLGELMEQG FVAPIPAGYQLSANANAVLA  140
AncOCPall         MSFTIESAQNIFSNTQVADA VPATIALFDQLSVDDQLALL WYAYTEMGRTITPAAPGAAS LQLAEGLLNQIKQMSPEEQT QVMRDLANHADTPISRSYGY FSVNTKLGFWYQLGEWMKQG IVAPIPAGYQMSPEVKAVLE  140
AncOCP1&2         MSFTIESARSIFPNTQVADA VPATIAAFNQLSAEDQLALL WFAYTEMGRTITPAAPGAAS MVLAEGTLNQIKQMSALEQT QVMCDLANHADTPICRTYGS FSTNIKLGFWYQLGEWMKQG IVAPIPEGYKLSADASAVLQ  140
AncOCP1           MPFTIESARGIFPGTLAADA VPATIARFNQLSAEDQLALI WFAYTEMGKTITIAAPGAAS MVLAEGTLNQIRQMSALEQT QVMCDLANRADTPICRTYGT FSPNIKLGFWYQLGEWMEQG IVAPIPEGYKLSANASAVLQ  140
conAncOCP1        MPFTIDSARGIFPNTLAADA VPATIARFNQLSAEDQLALI WFAYLEMGKTITIAAPGAAS MQFAEGTLNQIRQMSFQEQT QVMCDLANRADTPICRTYAT WSPNIKLGFWYQLGEWMEQG IVAPIPEGYKLSANASAVLQ  140
altAllAncOCPall   MTYTIESARSIFSGTQVPSP IPATIAMFDQLSVDDRLALL WYAYTEMGRTITPAAPGAAS MQLAEGILNQIKQMSPEEQT QVMRDLASRADTPISRSYGY FSVNTKLGFWYELGEWMKQG IVAPIPTGYQMSPEVKVVLE  140
altAncOCP1&2      MSFTIESARSIFPNTQVADA VPATVAAFNQLSAEDQLALL WFAYTEMGKTITPAAPGAAS MVLAEGLLNQIKQMSALEQT QVMCDLANHADTPICRTYGS FSVNVKLGFWYQLGEWMKQG IVAPIPEGYKLSANASAVLE  140
altAncOCP1        MPFTIESARGIFPNTLAADA VPATVARFNQLSAEDQLALI WFAYTEMGKTITIAAPGAAS MVLAEGLLNQIRQMSALEQT QVMCDLANRADTPICRTYAT FSPNIKLGFWYQLGQWMEQG IVAPIPEGYKLSANASAVLQ  140
altConAncOCP1     MPFTIDSARGIFPNTLSADA VPATIARFNQLSAEDQLALI WFAYLEMGKTITIAAPGAAS MQFAEGTLNQIRQMSFQEQT QVMCDLANRADTPICRTYAT WSPNIKLGFWYQLGQWMEQG IVAPIPEGYKLSANASAVLQ  140

                                                  LINKER                                 CTD
OCP1(SYNY3)       TIQGLESGQQITVLRNAVVD MGFT       AGKDG----KRIA-EPVVPPQDTASRTKVS         IEGVTN ATVLNYMDNLNANDFDTLIE LFTSDGALQPPFQRPIVGKE NVLRFFREECQNLKLIPERG  260
AncOCPall         AIKKLDPGQQITVLRNTVVN MGFD       PSDVASSSSPKKAAEPTFPRTAPAP-TKIT         IEGINE PTVLSYIEAMNADNFDAAVA LFAPDGALQPPFQKPIVGRE AIAAYMREEAQGLKMMPQQG  260
AncOCP1&2         AIRQLDAGQQITVLRNTVVN MGYD       PSKLG--GYKKVA-EPVVPPKDLAPRTKVS         IEGIDN PTVLSYIDNMNAFDFDAAVA LFAEDGALQPPFQKPIVGKE AILAYMREECQGLKLMPEQG  260
AncOCP1           AIRELDAGQQITVLRNAVVD MGYD       PSKLG--GYTRVA-EPVVPPKEMAQRTQVS         IEGVDN PTVLSYINNMNANDFDAVIN LFTGDGALQPPFQRPIVGKE AVLRYLREECQNLKLMPERG  260
conAncOCP1        AIRELDSGQQITVLRNSVVD MGFD       PNKLG--SYTRVS-EPVVPPKEMSQRTQVT         IEGVDN PTVLSYMNNMNANDFDALIE LFTPDGALQPPFQRPIVGKE AVLRFFREECQNLKLIPERG  260
altAllAncOCPall   AIKKLDAGQQITVLRNTVVD MGFD       PADLAPNSYAKKAAEPTFPRTELSP-TKVK         IEGITE PTVLSYIEAMNADDFEAAVA LFAPDGALQPPFQKPIVGRE AIATYMREEAQGLNMMPQQG  260
altAncOCP1&2      TIRELDAGQQITVLRNTVVN MGYD       PTRSG--GYKKVE-EPVVPPKEISQRTKVS         IEGIDN PTVLSYIENMNAFDFEAAVA LFAEDGALQPPFQKPIVGKE AILTYMREECQGLKLIPERG  260
altAncOCP1        TIRELDSGQQITVLRNAVVD MGYD       PNKLG--GYTRVS-EPVVPPKEVSQRTQVT         IEGVDN PTVLSYINNMNANDFDAVIN LFTPDGALQPPFQRPIVGKE AVLRYLREECQNLKLIPERG  260
altConAncOCP1     AIRELDSGQQITVLRNSVVD MGFD       PNKLG--SYTRVS-EPVVPPKEMSQRTQVT         IEGVDN PTVLSYMNNMNANDFDALIE LFTPDGALQPPFQRPIVGKE AVLRFFREECQNLKLIPERG  260

OCP1(SYNY3)       VTEPAEDGFTQIKVTGKVQT PWFGGNVGMNIAWRFLLNPE GKIFFVAIDLLASPKELLNF AR-  322;    extant protein length: 317aa; extant OCP1 of Synechocystis sp. PCC 6803
AncOCPall         ISETLEDGYTQVKVTGKVQT PWFGVNVGMNIAWRFLLNPQ GKIFFVAIDLLASPQELLNL RRR  322; ancestral protein length: 322aa; last common ancestor of all OCP
AncOCP1&2         VSEPAEDGYTQIKVTGKVQT PWFGANVGMNIAWRFLLNPQ GKIFFVAIDLLASPKELLNL GR-  322; ancestral protein length: 319aa; last common ancestor of OCP1 & OCP2
AncOCP1           VSEPAEDGYTQIKVTGKVQT PWFGANVGMNIAWRFLLNPQ GKIFFVAIDLLASPKELLNL VR-  322; ancestral protein length: 319aa; last common ancestor of OCP1
conAncOCP1        VSEPAEDGYTQIKVTGKVQT PWFGAGVGMNMAWRFLLNPE NKIFFVAIDLLASPKELLNL VR-  322; ancestral protein length: 319aa; conservative node sequence for LCA of OCP1
altAllAncOCPall   ISESLEDGYTQIKVTGKVQT PWFGANVGMNIAWRFLLNPQ GKIFFVAIDLLASPKELLNL RRK  322; ancestral protein length: 322aa; altAll sequence
altAncOCP1&2      VSEPAEDGFTQVKVTGKVQT PWFGANVGMNIAWRFLLNPQ GKIFFVAIDLLASPKELLNL RRK  322; ancestral protein length: 320aa; altAll sequence
altAncOCP1        VSEPAEDGFTQVKVTGKVQT PWFGAKVGMNIAWRFLLNPQ GKIFFVAIDLLASPKELLNL IR-  322; ancestral protein length: 319aa; altAll sequence
altConAncOCP1     VSEPAEDGYTQIKVTGKVQT PWFGAGVGMNMAWRFLLNPE NKIFFVAIDLLASPKELLNL AR-  322; ancestral protein length: 319aa; altAll sequence
```

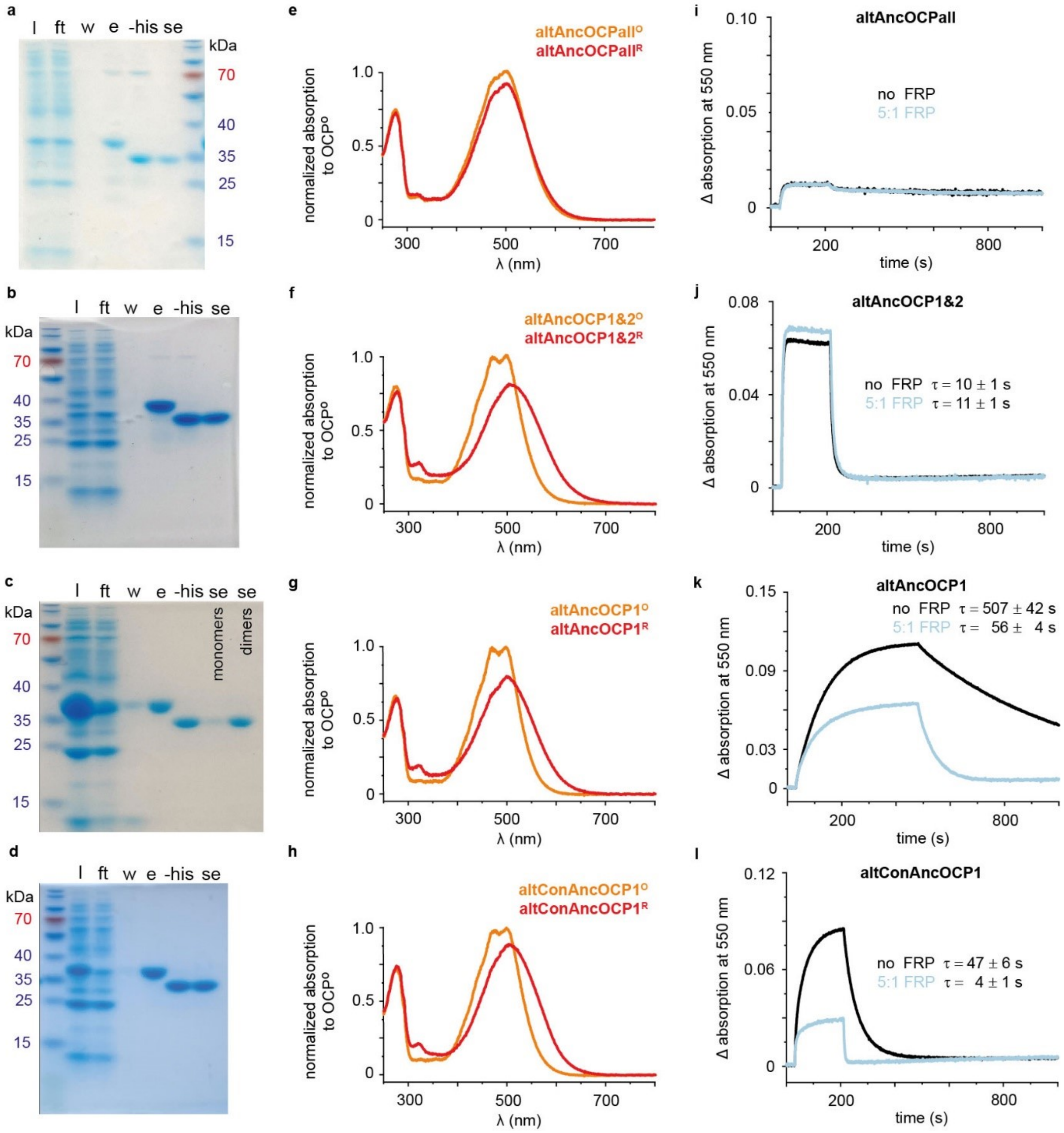
Extended Data Fig. 3| Reconstructed ancestral OCP sequences and their statistical robustness. a, Multiple sequence alignment of OCP1 from *Synechocystis* sp. PCC 6803 (SYNY3) with reconstructed ancestral OCP sequences and respective alternative sequences (alt). Important states for dimerization of OCP1^{O} [16], OCP1^{R} [29], deceleration of OCP1, and interaction with FRP[7] are indicated, and red if conserved or blue if not. Numeration of residues follows SYNY3 OCP1. C-terminal domain (CTD), linker, and N-terminal domain (NTD) regions are labelled accordingly. The more conservative ancestral OCP1 (conAncOCP1) and its alternative sequence that do not appear in the main text are greyed. **b-e**, Distribution of posterior probabilities (pp) per site with 20 bin categories per reconstructed sequence with the mean and the number of ambiguous sites shown. Sites were considered ambiguous if pp > 0.2 for the state with the second highest pp and were replaced with those states in the alt ancestors.

Extended Data Fig. 4| See following page for figure. Caption is below.

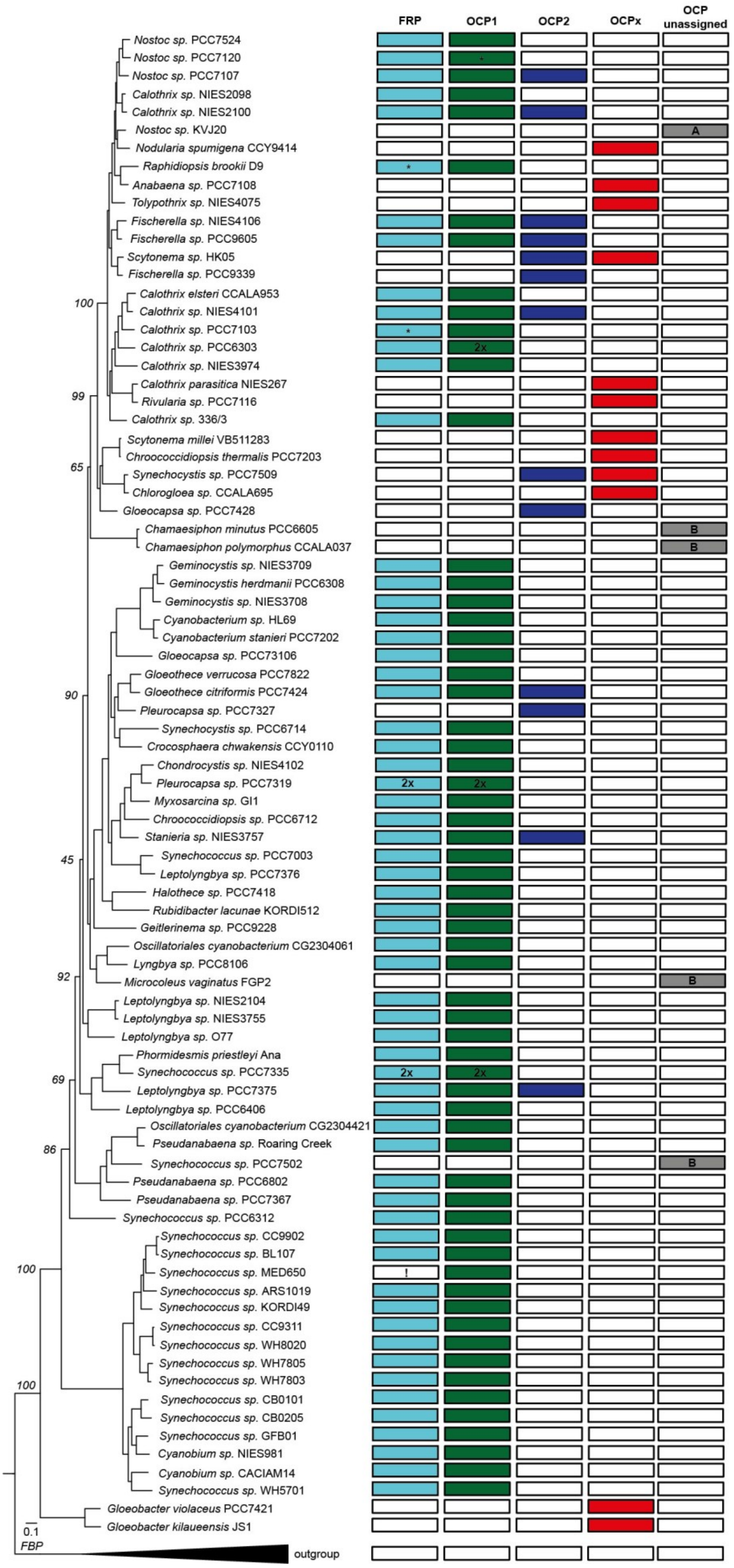
Extended Data Fig. 4| Biochemistry of ancestral OCPs. a-d, 12% SDS polyacrylamide gels of ancestral protein purifications. l, lysate. ft, flow through. w, wash. e, elution. -his, after his-tag cleavage. se, after size exclusion chromatography. Purifications were repeated three times with similar results. **e-h**, UV-Vis absorption spectra of inactive orange and active red state of ancestral OCPs. **i-p**, Recovery from photoconversion of ancestral OCPs with (in molar ratios of 5 OCP to 1 FRP) or without extant FRP from *Synechocystis* sp. PCC 6803 (SYNY3) as indicated at different temperatures. **q-t**, Arrhenius plots of recovery from photoconversion with (red) or without SYNY3 FRP (black). **u-y**, Recovery from photoconversion of ancestral OCPs either alone or with different ancestral FRPs or ancestral FRPLs or extant FRPL from *Pseudomonas borbori* in different molar ratios as indicated at 20 °C with respective mean recovery time constants (τ) and s.d. of three independent replicates. Representative data sets are shown for clarity.



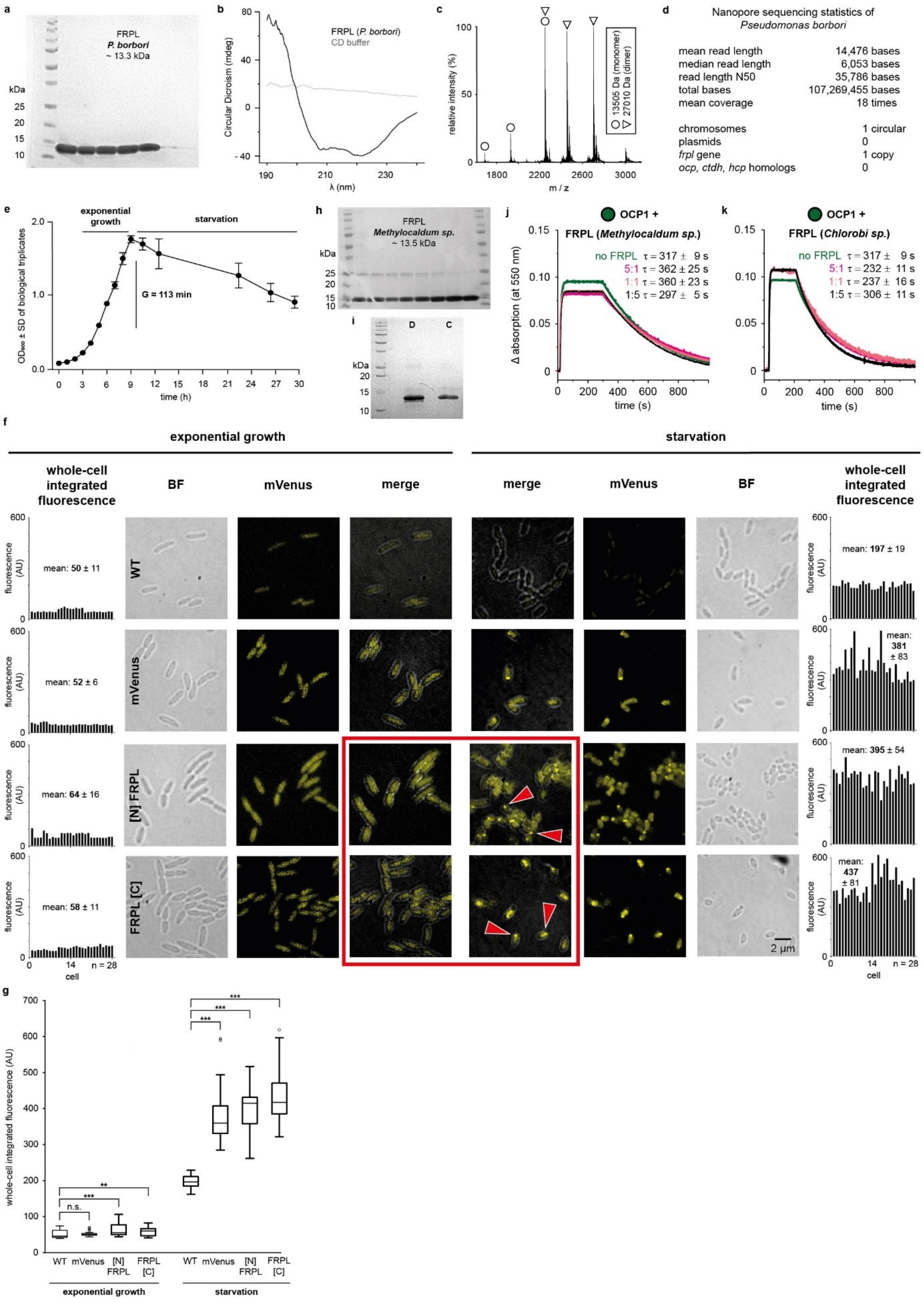
Extended Data Fig. 4| See previous page for caption.



Extended Data Fig. 5| Biochemistry of alternative ancestral OCPs. a-d, 12% SDS polyacrylamide gels of alternative ancestral protein purifications. l, lysate. ft, flow through. w, wash. e, elution. -his, after his-tag cleavage. se, after size exclusion chromatography. Purifications were repeated three times with similar results. **e-h**, UV-Vis absorption spectra of inactive orange and active red state of alternative ancestral OCPs. **i-l**, Recovery from photoconversion of alternative ancestral OCPs with (cyan) or without (black) extant FRP from *Synechocystis* sp. PCC 6803 at 20 °C with respective mean recovery time constants (τ) and s.d. of three independent replicates. Representative data sets are shown for clarity; altAncOCPall is barely photo-switchable.

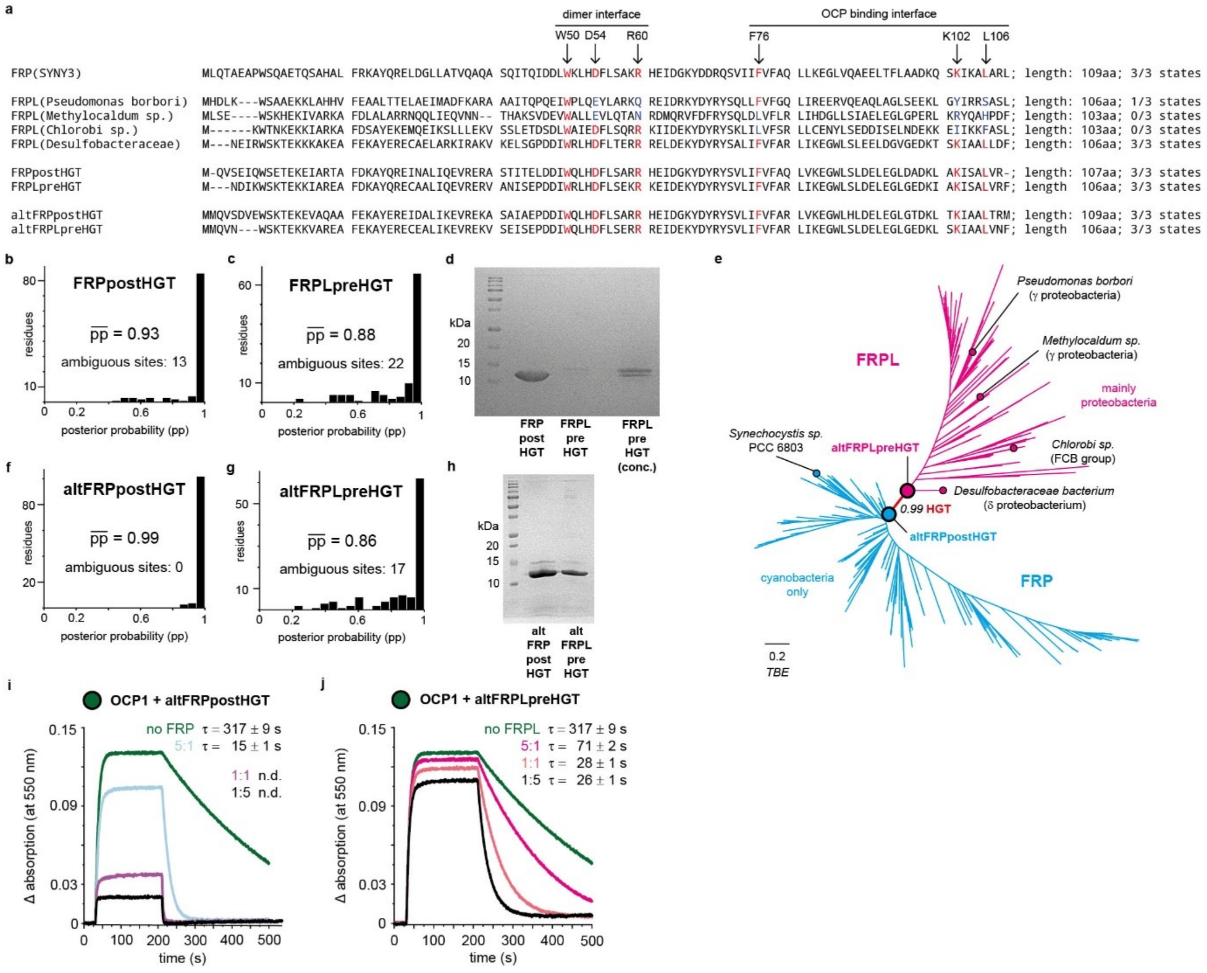


Extended Data Fig. 6| Species phylogeny of OCP-containing cyanobacteria. ML species phylogeny with Felsenstein Bootstrap Probabilities (FBP) of 100 replicates in italics. The appearance of FRP and OCP paralogs are mapped next to the phylogeny. Asterisks indicate multispecies entries in the BLAST database[39]. The exclamation point marks the only strain lacking FRP while having OCP1. Underlying amino acid sequence alignment in Supplementary Data 1.



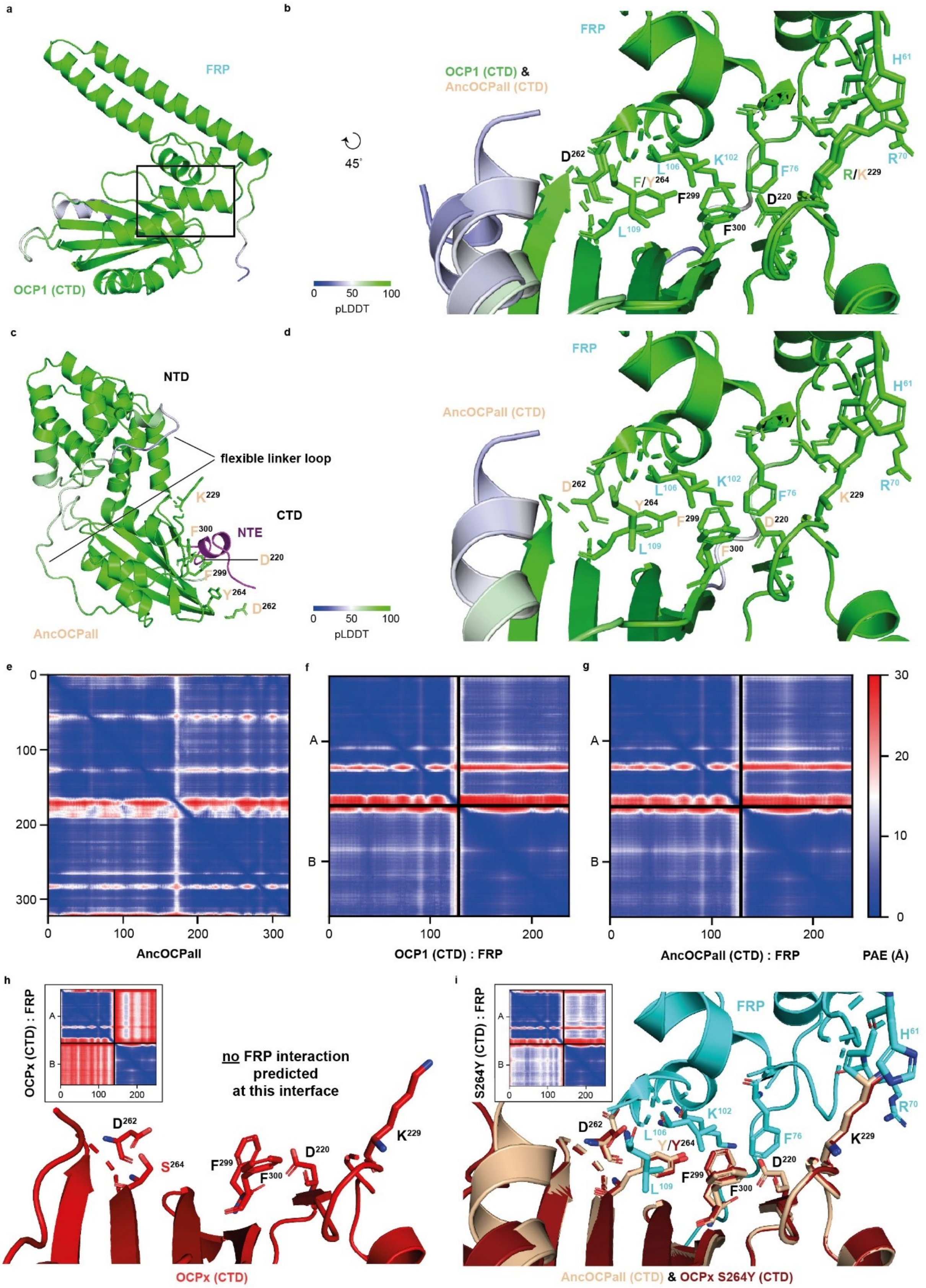
Extended Data Fig. 7| See next page for caption.

Extended Data Fig. 7| Characterization of extant FRPLs. a, h+i, 15% SDS polyacrylamide gels of *P. borbori*, *Methylocaldum* sp., *Desulfobacteriaceae* (D), and *Chlorobi* sp. (C) FRPL after size exclusion chromatography. Purifications were repeated three times with similar results. **b**, Circular dichroism (CD) spectra of *P. borbori* FRPL (black) in CD buffer (grey). **c**, Native mass spectrometry data of *P. borbori* FRPL. **d**, Nanopore sequencing statistics. **e**, Growth curve of *P. borbori* in biological triplicates with means and standard deviation (SD) shown, and determination of the generation time (G) during exponential growth. **f**, Epi-fluorescence microscopy of *P. borbori* strains expressing either none (WT), mVenus only (mVenus), or FRPL fusion proteins with either N- or C-terminal mVenus fusion. Whole-cell integrated fluorescence with the mean and SD, the brightfield (BF) image, the GFP channel signal (mVenus), and an overlay of both (merge) is shown. Red arrows point to signal foci at the cell poles. Scale bar represents 2 µm and is applicable for all images. **g**, Two-sided Welch's t-tests were performed to compare mean whole-cell integrated fluorescence with *** $p < 0.001$, ** $p = 0.013$, n.s., not significant ($p = 0.580$); $n = 28$ cells per condition. Boxes extend from lower to upper interquartile values of the data, with a line at the median. Whiskers display data within ± 1.5 interquartile ranges. Circles are outliers. **j+k,** Recovery from photoconversion of OCP1 from *Synechocystis* sp. PCC 6803 with extant FRPL as indicated at 20 °C with respective mean recovery time constants (τ) and SD of three independent replicates. Representative data sets are shown for clarity.



```
                                                                                dimer interface                    OCP binding interface
                                                                                W50 D54    R60                F76                            K102  L106
FRP(SYNY3)                       MLQTAEAPWSQAETQSAHAL FRKAYQRELDGLLATVQAQA SQITQIDDLWKLHDFLSAKR HEIDGKYDDRQSVIIFVFAQ LLKEGLVQAEELTFLAADKQ SKIKALARL; length: 109aa; 3/3 states

FRPL(Pseudomonas borbori)        MHDLK---WSAAEKKLAHHV FEAALTTELAEIMADFKARA AAITQPQEIWPLQEYLARKQ REIDRKYDYRYSQLLFVFGQ LIREERVQEAQLAGLSEEKL GYIRRSASL; length: 106aa; 1/3 states
FRPL(Methylocaldum sp.)          MLSE----WSKHEKIVARKA FDLALARRNQQLIEQVNN-- THAKSVDEVWALLEVLQTAN RDMQRVFDFRYSQLDLVFLR LIHDGLLSIAELEGLGPERL KRYQAHPDF; length: 103aa; 0/3 states
FRPL(Chlorobi sp.)               M------KWTNKEKKIARKA FDSAYEKEMQEIKSLLLEKV SSLETDSDLWAIEDFLSQRR KIIDEKYDYRYSKLILVFSR LLCENYLSEDDISELNDEKK EIIKKFASL; length: 103aa; 0/3 states
FRPL(Desulfobacteraceae)         M---NEIRWSKTEKKKAKEA FEKAYERECAELARKIRAKV KELSGPDDIWRLHDFLTERR RELDEKYDYRYSALIFVFAR LIKEGWLSLEELDGVGEDKT SKIAALLDF; length: 106aa; 3/3 states

FRPpostHGT                       M-QVSEIQWSETEKEIARTA FDKAYQREINALIQEVRERA STITELDDIWQLHDFLSARR HEIDGKYDYRYSVLIFVFAQ LVKEGWLSLDELEGLDADKL AKISALVR-; length: 107aa; 3/3 states
FRPLpreHGT                       M---NDIKWSKTEKKIAREA FDKAYQRECAALIQEVRERV ANISEPDDIWRLHDFLSEKR KEIDEKYDYRYSVLIFVFAR LIKEGWLSLDELEGLGEDKI AKISALVRF; length  106aa; 3/3 states

altFRPpostHGT                    MMQVSDVEWSKTEKEVAQAA FEKAYEREIDALIKEVREKA SAIAEPDDIWQLHDFLSARR HEIDGKYDYRYSVLIFVFAR LVKEGWLHLDELEGLGTDKL TKIAALTRM; length: 109aa; 3/3 states
altFRPLpreHGT                    MMQVN---WSKTEKKVAREA FEKAYERECEALIKEVREKV SEISEPDDIWQLHDFLSERR REIDEKYDYRYSVLIFVFAR LIKEGWLSLDELEGLGEDKL SKIAALVNF; length  106aa; 3/3 states
```

Extended Data Fig. 8| Resurrected ancestral FRP and ancestral FRPL sequences and their statistical robustness. a, Amino acid sequence alignment of FRP from *Synechocystis* sp. PCC 6803 (SYNY3) with extant and reconstructed ancestral FRPLs and ancestral FRPs. Important sites for homo-dimerization and interaction with OCP1 in FRP are pointed out[7,8], and red if conserved or blue if not. Numeration follows SYNY3 FRP. ML trees for the reconstructions in Supplementary Fig. 1+2. **b+c, f+g**, Distribution of posterior probabilities (pp) per site with 20 bin categories per reconstructed sequence with the mean and the number of ambiguous sites with pp > 0.2 for the state with the second highest pp shown. **d+h**, 15% SDS polyacrylamide gels of ancestral proteins after size exclusion chromatography. Purifications were repeated three times with similar results. conc., concentrated. **e**, Unrooted initial FRP(L) phylogenetic tree used for reconstruction of alternative (alt) ancestors at indicated nodes. Branch-lengths represent average substitutions per site. Full tree in Supplementary Fig. 2. HGT, horizontal gene transfer. TBE, Transfer Bootstrap Expectation. **i+j**, Recovery from photoconversion of SYNY3 OCP1 with alternative ancestral FRP (altFRPpostHGT) or alternative ancestral FRPL (altFRPLpreHGT) as indicated at different molar ratios at 20 °C with respective mean recovery time constants (τ) and s.d. of three independent replicates. Representative data sets are shown for clarity. n.d., not determinable.



Extended Data Fig. 9| See next page for caption.

Extended Data Fig. 9| The OCP-FRP interaction is predicted with high confidence. a+b, Per-residue estimate of confidence (pLDDT) of AlphaFold2 models shown in Fig. 4d+e. **c,** Confidence of the predicted full-length AncOCPall with indicated residues in the C-terminal domain (CTD) involved in the predicted interaction with FRP from *Synechocystis* sp. PCC 6803 (SYNY3) that are blocked by the N-terminal extension (NTE, in magenta) in the compact, orange state of AncOCPall predicted here. NTD, N-terminal domain. **d**, Confidence of the modelled interaction between AncOCPall and SYNY3 FRP. **e-g**, Predicted aligned errors (PAE). **h+i**, AlphaFold2 models of OCPx's CTD from *Gloeobacter kilaueensis* JS1 do not predict an interaction with SYNY3 FRP at the expected interface (consistent with experimental data[30]), unless serine (S) at position 264 (SYNY3 numeration) is changed for tyrosine (Y), the ancestral state in AncOCPall that is further shown in overlay here. Inserts show PAEs.

Methods

Molecular phylogenetics and ancestral sequence reconstruction

To infer the phylogenetic tree of cyanobacterial OCP proteins, we used the OCP dataset of Muzzopappa *et al.*[16], and profile-aligned the corresponding amino acid sequences of the three described OCP types therein (OCP1, OCP2, OCPx), using MUSCLE (v.3.8.31)[34]. We added sequences of either cyanobacterial CTD-like homologue proteins (CTDHs) or cyanobacterial HCPs as the respective outgroup. Alignments were corrected manually, sites corresponding to linage-specific insertions and duplicated sequences were removed. Full alignments are in Supplementary Data 1. We used RaxmlHPC-AVX (v.8.2.10)[35] in the PROTGAMMAAUTO mode to identify the best-fit model of amino acid evolution, which was the Revised Jones–Taylor–Thornton substitution matrix (JTTDCMut)[36] with empirical base frequencies and gamma distribution of among site rate-variation. We used PhyML (v.3.1)[37] with SPR moves to infer two ML phylogenies with either CTDH or HCP sequences included, and rooted the trees between either of those sequences and all OCP sequences on our trees. The two phylogenies show basically the same topology, but *unassigned grade A* is first branching on the HCP outgroup tree (Extended Data Fig. 2). As Gloeobacteria, which are known to be early branching cyanobacteria[18–21], only feature OCPx, but no OCP homologues of the unassigned grades, we used the CTDH outgroup tree for further analyses (Extended Data Fig. 1). The robustness of each topology was tested by running 100 non-parametric bootstraps, and additionally calculating aLRT statistics with PhyML. The ancestral OCP sequences were reconstructed at the internal node on the CTDH outgroup tree, as indicated in Fig. 1b and Extended Data Fig. 1, using marginal reconstruction in the CodeML module of PAML (v.4.9)[38] with the JTTDCMut substitution model and 16 gamma categories. Ancestral sequences were cropped following parsimony rules and contain the states with the highest posterior probabilities (pp) at all sites selected. The average pp values for all reconstructed proteins are in Extended Data Fig. 3b–e. The 'altAll' alternative sequences for every reconstructed ancestor comprises the state with the second highest pp if that state has pp > 0.20, and the ML state otherwise.

For the FRP(L) phylogenetic tree (Fig. 2a), we gathered amino acid sequences using online BLASTP[39] on 23 February 2022, and the FRP amino acid sequence of *Synechocystis* sp. PCC 6803 (SYNY3) as a query. To specifically find FRPL sequences, we excluded cyanobacteria (taxid:1117) and repeated the search against SYNY3 FRP and subsequently against *P. borbori* FRPL or explicitly searched in taxonomic groups other than cyanobacteria. Additionally, we added metagenomic sequences from the Global Microbial Gene Catalog (GMGC, v.1.0)[40]. Sequences were aligned with MUSCLE (v.3.8.31). The alignment was corrected manually, sites corresponding to linage-specific insertions and duplicated sequences were removed. The full alignment is in Supplementary Data 1. We used RaxmlHPC-AVX (v.8.2.10) in the PROTGAMMAAUTO mode using the Akaike information criterion to identify the best-fit model of amino acid evolution, which was the Le-Gascuel substitution matrix[41] with fixed base frequencies and gamma distribution of among site rate-variation. We inferred the ML phylogeny, and tested the robustness of the topology by running 100 non-parametric bootstraps. TBEs were calculated with the BOOSTER web tool[42]. Furthermore, aLRT statistics were calculated with PhyML (v.3.1). The tree was rooted between acidobacteria and proteobacteria in the FRPL group and suggests a HGT from an ancestral delta-proteobacterium into an ancestral cyanobacterium. The full tree is in Supplementary Fig. 1. Ancestral FRPL and ancestral FRP sequences (FRPLpreHGT and FRPpostHGT, respectively) were reconstructed at the internal nodes of the tree using marginal reconstruction in the CodeML module of PAML (v.4.9) with the Le-Gascuel substitution matrix (LG) model and 16 gamma categories. Gaps were assigned using parsimony. For the ancestors we resurrected, we chose the amino acid state with the highest pp at each site. The average pp for the reconstructed proteins are in Extended Data Fig. 8b,c.

For the gene tree–species tree reconciliation, we identified all sequences on our FRP(L) tree that could certainly be assigned to a distinct bacterial strain that is also deposited at the Genome Taxonomy Database (GTDB)[43] with its set of 120 single copy marker protein sequences, using BLASTP[39]. With these aligned, concatenated amino acid sequences, we inferred a ML phylogenetic tree using IQ-Tree 2 (v.2.2)[44] (-m LG, -b 100, -alrt 1,000), and rooted with acidobacteria as described above. We accordingly inferred a gene tree with FRP and FRPL sequences of the corresponding species, and ran 100 non-parametric bootstraps for this subset. Reconciliation was performed using ML estimation with *ALEml_undated* in ALE[45] and the rooted species phylogeny as well as the FRP(L) bootstrap trees as the input. Reconciled trees and ALE output are deposited in the source data.

To reconstruct the alternative ancestral FRPL and alternative ancestral FRP sequences (altFRPLpreHGT and altFRPpostHGT, respectively), we used an initial alignment with fewer sequences in total. The full alignment is in Supplementary Data 1. An ML phylogenetic tree with 100 non-parametric bootstraps was inferred, and the alternative ancestral FRPL and alternative ancestral FRP sequences were reconstructed accordingly at the internal node of that tree, shown in Extended Data Fig. 8e and

Supplementary Fig. 2, using marginal reconstruction in the CodeML module of PAML (v.4.9) with the Le-Gascuel substitution matrix substitution model and 16 gamma categories. TBE were calculated with the BOOSTER web tool. Alternative ancestral sequences were cropped following parsimony rules and contain the states with the highest pp at all sides selected. The average pp for the reconstructed proteins are in Extended Data Fig. 8f,g.

For the phylogenetic species tree of OCP-containing cyanobacteria, we identified all sequences on our OCP tree that could certainly be assigned to a distinct cyanobacterial strain that is also deposited at the GTDB with its set of 120 single copy marker protein sequences. As an outgroup, we added sequence sets of closely related malainabacteria as well as sets of more distantly related *Chloroflexota* species. We used these concatenated amino acid sequences, aligned them, and inferred a phylogenetic tree using RaxmlHPC-AVX (v.8.2.10) in the PROTGAMMAAUTO mode, using the Akaike information criterion to identify the best-fit model of amino acid evolution, which was the Le-Gascuel substitution matrix[41] with empirical base frequencies and gamma distribution of among site rate-variation. We inferred the ML phylogeny, and tested the robustness of the topology by running 100 non-parametric bootstraps. We rooted the tree between cyanobacteria and the outgroup, and mapped the appearance of *frp* and *ocp* genes in corresponding genomes, on the basis of BLASTP and tBLASTn[39] hits, next to the tree (Extended Data Fig. 6). Assignment of particular OCP sequences to an OCP paralogue group is based on the position of their translated amino acid sequences on our OCP tree (Extended Data Fig. 1).

Cloning and protein purification

DNA sequences of ancestral OCPs, extant OCP1 from *Synechocystis* sp. PCC 6803 (SYNY3) and FRP (SYNY3) were codon optimized for expression in *E. coli*, and synthesized by either Genscript Biotech or Life Technologies (GeneArt). Synthesized constructs were flanked by *BamHI* and *NotI* cleaving sites for cloning into a modified pRSFDuet-1 vector (Merck Millipore), which encodes a specific human rhinovirus HRV 3 C protease cleavage site (LEVLFQ/GP) and a 6xHis tag at the N terminus (resulting plasmid termed pRSFDuetM). After cleavage, all constructs started with GPDPATM. For expression of extant FRP (SYNY3 gene *slr1964*), the pRSFDuetM-FRP vector was transformed into *E. coli* BL21 (DE3) (New England Biolabs), which were grown overnight at 37 °C in Luria–Bertani (LB) medium (1% tryptone, 1% NaCl, 0.5% yeast extract, pH 7.0), supplemented with kanamycin (Kan, 50 µg ml^{-1}). The following day, 1 l of LB + Kan was inoculated with 10 ml of overnight culture, and incubated at 37 °C until an optical density (OD_{600nm}) of 0.6–0.8, then induced by 0.5 mM isopropyl-β-d-thiogalactopyranoside (IPTG) and grown in a shaking incubator for 24 h at 30 °C. Cells were gathered at 10,000 g for 10 min, and stored at −20 °C until use. For expression of OCPs (extant OCP1, SYNY3 gene *slr1963* and ancestral OCPs), the corresponding pRSFDuetM-OCPxx constructs were transformed into echinenone-producing *E. coli* BL21 (DE3), harboring a p25crtO plasmid. The expressions were carried out in 1 l of LB, supplemented with chloramphenicol (34 µg ml^{-1}) and Kan (50 µg ml^{-1}), which was inoculated by 10 ml of overnight culture, and grown in a shaking incubator at 37 °C until OD_{600nm} = 0.6–0.8. After induction with 0.5 mM IPTG, cells were incubated at 25 °C for 72 h, and finally collected at 10,000 g for 10 min and stored at −20 °C until use.

For purification, frozen cell pellets were resuspended in phosphate-buffered saline (PBS) (137 mM NaCl, 2.7 mM KCl, 12 mM phosphate, pH 7.4), supplemented with 100 mg of lysozyme (Ovobest) and protease inhibitor (1 mM benzamidine, 1 mM ε-amino-caproic acid). Cell lysis was performed by using a FrenchPress (G. Heinemann) in three cycles at 18,000 psi. Afterwards, cell debris was pelleted at 18,000 g for 15 min at 4 °C. Supernatant was loaded on a 5 ml Co^{2+}-HiTrap Talon crude column (Cytiva) using a peristaltic pump. Elution was carried out with imidazole-containing buffer (1×PBS + 350 mM imidazole, pH 7.4), supplemented with HRV 3C protease in a total mass ratio of 500:1 (protein to protease) and dialyzed at 4 °C in 3C protease buffer (20 mM Tris, 100 mM NaCl, 2 mM dithiothreitol, pH 8.5) for 18 h. Protein solution was reloaded on a Co^{2+}-HiTrap Talon crude column while this time, flow through was collected. In case of FRP, purification was performed by SEC for polishing, while OCP purification was continued with hydrophobic interaction chromatography (HIC) to remove apo-protein. Collected OCP flow-throughs were dialyzed overnight in HIC buffer (500 mM $(NH_4)_2SO_4$, 100 mM urea, 5 mM phosphate, pH 7.5) at 4 °C. HIC was performed on a HiPrepTM 16/10 Phenyl HP column (Cytiva) in an automated Azura FPLC system (Knauer). Proteins were eluted with a hydrophilic buffer (100 mM urea, 5 mM phosphate, pH 7.5). Carotenoid-rich protein fractions were concentrated using 10 kDa molecular weight cut-off (MWCO) centrifugal filter units (Pall Corporation) for SEC. FRP was concentrated with 3 kDa MWCO centrifugal filter units. Then, 500 µl of each concentrated protein solutions were loaded on a SuperdexTM 200 Increase 10/300 column (Cytiva) and eluted with 1× PBS. Proteins were stored at −80 °C until use.

Codon-optimized sequences coding for extant FRPL, ancestral FRP, and ancestral FRPL proteins were obtained from Integrated DNA Technologies (IDT) or Twist Biosciences. They were cloned into pET-LIC vectors containing an N- or C-terminal 6xHis tag using Gibson Assembly Master Mix (New England Biolabs). The oligonucleotides used are shown in Supplementary Tab. 1. Correct assembly was verified by Sanger Sequencing (Microsynth). Plasmids were transformed into *E. coli* BL21 (DE3) (Invitrogen). For protein overproduction, 50 ml of LB, supplemented with carbenicillin (Carb) (100 µg ml^{-1}), were inoculated with a single colony from a fresh LB + Carb plate, and grown overnight at 37 °C in a shaking incubator. Six lots of 500 ml of LB + Carb were inoculated with overnight cultures at OD_{600nm} = 0.01, and grown to OD_{600nm} = 0.6–0.8 for roughly 2.5 h. Protein overproduction was induced with 1 mM IPTG. After 4 h, cells were gathered at 4,392 g for 20 min at 4 °C and cell pellets were stored at −20 °C until usage.

For purification, cells were resuspended in 35 ml of buffer A (300 mM NaCl, 20 mM Tris, 20 mM imidazole, 5 mM β-mercaptoethanol, pH 8.0), and one tablet of cOmplete Protease Inhibitor Cocktail (Roche) was added. Cells were disrupted twice in an LM10 microfluidizer (Microfluidics) at 13,000 psi. Lysate was cleared by centrifugation at 29,930 g for 30 min, and being passed through a 0.45 µm syringe filter, then loaded on a 5 ml Bio-Scale Mini Nuvia Ni-charged IMAC Cartridge (BioRad). After washing with 25 ml of buffer A, protein was eluted with a linear gradient over 20 ml from 0 to 100% of buffer B (300 mM NaCl, 20 mM Tris, 500 mM imidazole, 5 mM β-mercaptoethanol, pH 8.0) in an NGC system (BioRad). Fractions containing the protein were verified on in-house casted 15% SDS gels, and were pooled for SEC with a HiLoad 26/600 Superdex column (Cytiva) in SEC buffer (200 mM NaCl, 20 mM KCl, 20 mM HEPES, pH 7.5) in an NGC system. Purity of the fractions containing the protein were verified on in-house casted 15% SDS gels, and were pooled for concentration at 2,000 g with Amicon Ultra centrifugal filter units (Millipore) with a MWCO of 3 kDa. Proteins were stored at −20 °C until usage.

Carotenoid extraction and ultra-fast liquid chromatography analysis.
To analyze the carotenoid content of OCP holo-proteins, 50 µl of concentrated protein solution was mixed with 1 ml of acetone and centrifuged at maximum speed at 4 °C to spin down precipitated protein. Yellowish supernatant was evaporated in a centrifugal vacuum concentrator (Eppendorf) at 30 °C until the acetone evaporated completely and carotenoids had precipitated as red crystals. Remaining water solution was removed, and red carotenoid crystals were re-dissolved in 50 µl of acetone. The carotenoid-rich solution was transferred into a sample vial that was placed in an UFLC NexeraX2 system (Shimadzu), equipped with an Accucore C30 column (Thermo Fisher Scientific, 250 × 2.1 mm, 2.6 µm particle size, 150 Å pore size). As mobile phase eluents, buffer A (methanol to water, 95:5) and buffer B (methanol to THF, 7:3) were used with the following protocol: 0–4.3 min 0% of buffer B, 4.3–8.6 min linear gradient from 0 to 100% of buffer B, 8.6–15.6 min 100% of buffer B, 15.6–20.1 min 0% of buffer B with a constant flow rate of 0.4 ml min^{-1}. Eluted carotenoids were verified by mass spectrometry to correlate elution times with specific carotenoid species as well as by thin-layer chromatography and comparison with reference samples.

Ultraviolet–visible spectroscopy and kinetic analysis
Absorption spectra were recorded with a Maya2000Pro spectrometer (Ocean Optics), coupled via a fibre to a deuterium tungsten light source (Sarspec) and a cuvette holder (CVH100, Thorlabs). For OCP/FRP kinetic analyses, a temperature-controlled cuvette holder with a constant stirring device *qpod2e* (Quantum Northwest) was fiber-coupled to a CCS100/M spectrometer (Thorlabs) and a SLS201L/M tungsten light source (Thorlabs). For illumination with actinic light, a 3 W light-emitting diode (Avonec) with a maximum emission at 455 nm was used. Different OCP^O (mixed with different extant or ancestral FRP or extant or ancestral FRPL in various molar ratios, or alone) were photo-switched into the red state (OCP^R) by applying blue light for at least 3 min and 30 s or until a plateau was reached, and photo-recovery was constantly followed at 550 nm after turning off the blue light source. Recovery time constants (τ) were determined by fitting relaxation curves of the OCP^R to OCP^O back-conversions with a mono-exponential decay function and standard deviations (s.d.) of three independent replicates were calculated.

Circular dichroism spectroscopy
Far-ultraviolet circular dichroism spectroscopy was used to assess the secondary structure of heterologously produced *P. borbori* FRPL (PbFRPL) in solution. The protein was diluted to a concentration of roughly 50 µg ml^{-1} in circular dichroism Buffer (100 mM NaF, 10 mM Na_2HPO_4/ NaH_2PO_4, pH 7.5), and was measured in a 0.1 cm cuvette at room temperature using a JASCO J-810 spectropolarimeter (Jacso) in the range of 190–240 nm in 0.2 nm scanning steps. Three successive spectra were recorded, baseline corrected and averaged.

Native mass spectrometry

FRPL protein sample from *P. borbori* (PbFRPL) was stored at −20 °C before being buffer exchanged into 200 mM ammonium acetate (pH 6.8) by multiple rounds of concentration and dilution using Pierce protein concentrators (Thermo Fisher Scientific). The sample was then diluted to 4 µM (monomer) immediately before the measurements. Data were collected using in-house gold-plated capillaries on a Q Exactive mass spectrometer (Thermo Fisher), operated in positive ion mode with a source temperature of 100 °C and a capillary voltage of 1.2 kV. In-source trapping was set to −100 V to help with the dissociation of small ion adducts. Ion transfer optics and voltage gradients throughout the instruments were optimized for ideal transmission. Spectra were acquired with ten micro-scans to increase the signal-to-noise ratio with transient times of 64 ms, corresponding to the resolution of 17,500 at m/z = 200, and AGC target of 1.0×10^6. The noise threshold parameter was set to three and the scan range used was 350 to 8,000 m/z.

X-ray crystallography

Crystallization of *P. borbori* FRPL (PbFRPL) was performed by the hanging-drop method at 20 °C in 2 µl drops, consisting of equal amounts of protein and precipitation solutions. PbFRPL crystallized at 119 µM within 20 days in 0.2 M Li_2SO_4, 0.1 M CHES, pH 9.5 and 1.4 M sodium-potassium tartrate. Before data collection, crystals were flash-frozen in liquid nitrogen without the use of cryo-protectants. Synchrotron data were collected under cryogenic conditions at the P13 beamline, operated by the European Molecular Biology Laboratory (EMBL) Hamburg at the PETRA III storage ring (Deutsches Elektronen Synchrotron)[46]. Data were integrated and scaled with XDS, and merged with XSCALE[47]. Structures were determined by molecular replacement with PHASER[48], manually built in COOT[49] and refined with PHENIX[50]. For structure determination by molecular replacement, the crystal structure of FRP from *Synechocystis* sp. PCC 6803 (PDB ID 4JDX[25]) was used as a search model. The final structure of PbFRPL was uploaded to the RCSB PDB under accession number 8AG8. Data were rendered and visualized with PyMol (v.2.4.0)[51].

Whole-genome nanopore sequencing

After several rounds of cultivation, we re-sequenced the whole genome of *P. borbori* to rule out *frpl* gene loss on cultivation (a possible explanation for absence of FRPL in all model organisms), plasmid localization (that could facilitate HGT) or sample contamination, but found the genome to be a single, circular chromosome of 5.34 MB in size, entailing one copy of the *frpl* gene, but no OCP, HCP or CTDH homologues (Extended Data Fig. 7d). Genomic DNA of stationary phase *P. borbori* was obtained using the NucleoBond HMW DNA kit (Macherey-Nagel) according to the manufacturer's guidelines, and using lysozyme for cell lysis (final concentration 1 mg ml^{-1}) for 1 h at 37 °C in 2 ml of 10 mM Tris-HCl, pH 8.0. DNA quality and concentration were assessed via NanoDrop 8000 spectrophotometer and Qubit 3 fluorometer using double-stranded DNA BR reagents. Library preparation was performed using the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies), according to the manufacturer's guidelines, except the input DNA was increased fivefold to match the molarity expected in the protocol as no DNA shearing was applied. Sequencing was performed on a MinION Mk1B device for 24 h using a 'Flongle Flow Cell' (FLO-FLG001, cell chemistry R9.4.1). Nanopore data were base-called with ONT Guppy base-calling software. Long reads were assembled using canu[52], resulting in a single circular chromosome. Raw reads are deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive and can be accessed under BioProject no. PRJNA865569 and BioSample accession no. SAMN30120905.

Cultivation and genetic manipulation of *P. borbori*

The type stain DSM17834 of the delta-proteobacterium *P. borbori* was purchased from the German Collection of Microorganisms and Cell Cultures (Braunschweig, Germany). It was cultivated aerobically in PME medium (0.5% peptone, 0.3% meat extract, pH 7.0) at 28 °C, and a growth curve of biological triplicates was recorded. The generation time (G) during exponential growth was estimated using the formula: $G = \frac{\Delta t}{3.3\, log\left(\frac{OD2}{OD_1}\right)}$.

Protein fusions for in vivo localization with epi-fluorescence microscopy were generated by PCR amplification of the *frpl* gene of *P. borbori* including 200 bp of the 5′ untranslated region and insertion into pSG1164 vectors with an N- or C-terminal mVenus coding sequence and a 'GGGGGSL' linker sequence in frame using Gibson Assembly Master Mix (NEB). Correct assembly was verified by Sanger Sequencing (Microsynth). Chemically competent *P. borbori* were prepared by modification of a protocol by Irani and John[53], initially developed for *P. aeruginosa*, as follows: the medium was changed to PME, and temperatures were lowered to 28 °C. Plasmids were transformed into *P. borbori* following the transformation protocol of Irani and John[53], but changing the heat shock temperature to 30 °C, the medium to PME, the growth temperature to 28 °C and the carbenicillin concentration to 100 µg ml^{-1}. Plates were incubated at 28 °C for 48 h until colonies were visible.

Epi-fluorescence microscopy

For epi-fluorescence microscopy, *P. borbori* cells were grown at 28 °C and 200 r.p.m. to OD_{600nm} = 0.6 for 'exponential growth' and for 2 days to OD_{600nm} of around 1.0 for 'starvation' conditions in PME media. Cells were fixed on 1% agarose pads by sandwiching 100 µl of melted agarose between two coverslips (12 mm, Menzel). Then 3 µl of the culture was added onto a round coverslip (25 mm; Marienfeld) and fixed with an agarose pad. For widefield image acquisition, a Zeiss Observer A1 microscope (Carl Zeiss) with an oil immersion objective (×100 magnification, 1.45 numerical aperture, alpha Plan-FLUAR; Carl Zeiss) was used with a charge-coupled-device camera (CoolSNAP EZ; Photometrics) and an HXP 120 metal halide fluorescence illumination with intensity control. For epi-fluorescence microscopy, a green fluorescent protein filter set was used (BrightLine 470/40, Beamsplitter 495 and Brightline 525/50). Samples were illuminated for 0.5 to 2 s at mid-cell plane. Whole-cell integrated fluorescence was determined per cell and corrected for background fluorescence. Final editing of images was done in ImageJ2/FIJI (v.1.52)[54,55].

Analytical Size Exclusion Chromatography

Analytical SEC was performed with a Superdex 75 Increase 3.2/300 column (Cytiva), equilibrated with 1× PBS at a flow rate of 0.1 ml min^{-1} and a total sample injection volume of 20 µl. For measuring at blue light illumination, four 3 W LEDs (Avonec) with an emission maximum at 455 nm were mounted on a 20 cm heat sink at constant distances in front of the SEC column to continuously illuminate the sample on the column. Absorption was recorded at 280, 496 and 550 nm to follow elution profiles.

AlphaFold2 protein complex prediction

AlphaFold2 protein complex models were generated using the ColabFold server[56] on 20 May 2022, using as input sequences the CTD of either OCP1 from *Synechocystis* sp. PCC 6803 (SYNY3) or AncOCPall and FRP (SYNY3) with default settings. Further, the structure of full-length AncOCPall was predicted separately. On 3 November 2022, we repeated the analysis with the CTD of OCPx from *G. kilaueensis* JS1 or an S264Y mutant (serine at position 264 (SYNY3 numeration) was changed to tyrosine) of that OCPx with FRP (SYNY3). Modelled structures are deposited in the source data. Data were rendered and visualized with PyMol (v.2.4.0)[51].

Native PAGE

Native PAGE was performed in a Mini-Protean Tetra Cell (Biorad) by using in-house casted gradient gels with 3–14% acrylamide concentration in a Tris-glycine buffer system without SDS to obtain native protein conditions. No stacking gel was used. The electrophoresis chamber was constantly cooled in a fridge and illuminated by four 3 W LEDs (Avonec) with an emission maximum at 455 nm to photo-switch the OCP proteins in-gel. The voltage was set to 80 V constantly for 240 min, and subsequently to 120 V for another 100 min.

References

1. Peracchi, A. & Mozzarelli, A. Exploring and exploiting allostery: models, evolution, and drug targeting. *Biochim. Biophys. Acta* **1814**, 922–933 (2011).

2. Dawkins, R. *Climbing Mount Improbable* (Norton, 1996).

3. Pillai, A.S. *et al.* Origin of complexity in haemoglobin evolution. *Nature* **581**, 480–485 (2020).

4. Coyle, S.M., Flores, J. & Lim, W.A. Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell* **154**, 875–887 (2013).

5. Bridgham, J.T., Carroll, S.M. & Thornton, J.W. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**, 97–101 (2006).

6. Pillai, A.S., Hochberg, G.K.A. & Thornton, J.W. Simple mechanisms for the evolution of protein complexity. *Protein Sci.* **31**, e4449 (2022).

7. Muzzopappa, F. & Kirilovsky, D. Changing color for photoprotection: the orange carotenoid protein. *Trends Plant Sci.* **25**, 92–104 (2020).

8. Slonimskiy, Y.B., Maksimov, E.G. & Sluchanko, N.N. Fluorescence recovery protein: a powerful yet underexplored regulator of photoprotection in cyanobacteria. *Photochem. Photobiol. Sci.* **19**, 763–775 (2020).

9. Kay Holt, T. & Krogmann, D.W. A carotenoid-protein from cyanobacteria. *Biochim. Biophys. Acta* **637**, 408–414 (1981).

10. Wilson, A. *et al.* A soluble carotenoid protein involved in phycobilisome-related energy dissipation in cyanobacteria. *Plant Cell* **18**, 992–1007 (2006).

11. Wilson, A. *et al.* A photoactive carotenoid protein acting as light intensity sensor. *Proc. Natl Acad. Sci. USA* **105**, 12075–12080 (2008).

12. Gwizdala, M., Wilson, A. & Kirilovsky, D. *In vitro* reconstitution of the cyanobacterial photoprotective mechanism mediated by the orange carotenoid protein in *Synechocystis* PCC 6803. *Plant Cell* **23**, 2631–2643 (2011).

13. Bao, H. *et al.* Additional families of orange carotenoid proteins in the photoprotective system of cyanobacteria. *Nat. Plants* **3**, 17089 (2017).

14. Boulay, C., Wilson, A., D'Haene, S. & Kirilovsky, D. Identification of a protein required for recovery of full antenna capacity in OCP-related photoprotective mechanism in cyanobacteria. *Proc. Natl Acad. Sci. USA* **107**, 11620–11625 (2010).

15. Thurotte, A. *et al.* The cyanobacterial fluorescence recovery protein has two distinct activities: orange carotenoid protein amino acids involved in FRP interaction. *Biochim. Biophys. Acta, Bioenerg.* **1858**, 308–317 (2017).

16. Muzzopappa, F., Wilson, A. & Kirilovsky, D. Interdomain interactions reveal the molecular evolution of the orange carotenoid protein. *Nat. Plants* **5**, 1076–1086 (2019).

17. Melnicki, M.R. *et al.* Structure, diversity, and evolution of a new family of soluble carotenoid-binding proteins in cyanobacteria. *Mol. Plant* **9**, 1379–1394 (2016).

18. Schirrmeister, B.E., Gugger, M. & Donoghue, P.C.J. Cyanobacteria and the great oxidation event: evidence from genes and fossils. *Palaeontology* **58**, 769–785 (2015).

19. Moya, A. *et al.* Driven progressive evolution of genome sequence complexity in cyanobacteria. *Sci. Rep.* **10**, 19073 (2020).

20. Moore, K.R. *et al.* An expanded ribosomal phylogeny of cyanobacteria supports a deep placement of plastids. *Front. Microbiol.* **10**, 1612 (2019).

21. Rahmatpour, N. *et al.* A novel thylakoid-less isolate fills a billion-year gap in the evolution of cyanobacteria. *Curr. Biol.* **31**, 2857–2867.e4 (2021).

22. Kirilovsky, D. & Kerfeld, C.A. The orange carotenoid protein: a blue-green light photoactive protein. *Photochem. Photobiol. Sci.* **12**, 1135–1143 (2013).

23. Coleman, G.A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).

24. Vanparys, B., Heylen, K., Lebbe, L. & de Vos, P. *Pseudomonas peli* sp. nov. and *Pseudomonas borbori* sp. nov., isolated from a nitrifying inoculum. *Int. J. Syst.* **56**, 1875–1881 (2006).

25. Sutter, M. *et al.* Crystal structure of the FRP and identification of the active site for modulation of OCP-mediated photoprotection in cyanobacteria. *Proc. Natl Acad. Sci. USA* **110**, 10022–10027 (2013).

26. Sluchanko, N.N., Slonimskiy, Y.B., Moldenhauer, M., Friedrich, T. & Maksimov, E.G. Deletion of the short N-terminal extension in OCP reveals the main site for FRP binding. *FEBS Lett.* **591**, 1667–1676 (2017).

27. Sluchanko, N.N. *et al.* OCP-FRP protein complex topologies suggest a mechanism for controlling high light tolerance in cyanobacteria. *Nat. Commun.* **9**, 3869 (2018).

28. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

29. Domínguez-Martín, M.A. *et al.* Structures of a phycobilisome in light-harvesting and photoprotected states. *Nature* **609**, 835–845 (2022).

30. Slonimskiy, Y.B. *et al*. A primordial orange carotenoid protein: structure, photoswitching activity and evolutionary aspects. *Int. J. Biol. Macromol.* **222**, 167–180 (2022).

31. Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).

32. Petrescu, D.I., Dilbeck, P.L. & Montgomery, B.L. Environmental tuning of homologs of the orange carotenoid protein-encoding gene in the cyanobacterium *Fremyella diplosiphon*. *Front. Microbiol.* **12**, 819604 (2021).

33. Schulz, L., Sendker, F.L. & Hochberg, G.K.A. Non-adaptive complexity and biochemical function. *Curr. Opin. Struct. Biol.* **73**, 102339 (2022).

34. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

35. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**,1312–1313 (2014).

36. Kosiol, C. & Goldman, N. Different versions of the Dayhoff rate matrix. *Mol. Biol. Evol.* **22**, 193–199 (2005).

37. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

38. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

39. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

40. Mende, D.R. *et al.* proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res.* **48**, D621–D625 (2020).

41. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

42. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).

43. Parks, D.H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

44. Minh, B.Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

45. Szöllősi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

46. Cianci, M. *et al.* P13, the EMBL macromolecular crystallography beamline at the low-emittance PETRA III ring for high- and low-energy phasing with variable beam focusing. *J. Synchrotron Rad.* **24**, 323–332 (2017).

47. Kabsch, W. XDS. *Acta Crystallogr. D.* **66**, 125–132 (2010).

48. McCoy, A.J*. et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

49. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D.* **60**, 2126–2132 (2004).

50. Adams, P.D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D.* **66**, 213–221 (2010).

51. The PyMOL Molecular Graphics System v.2.4.0 (Schrödinger, LLC).

52. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

53. Irani, V.R. & Rowe, J.J. Enhancement of transformation in *Pseudomonas aeruginosa* PAO1 by Mg^{2+} and heat. *BioTechniques* **22**, 54–56 (1997).

54. Rueden, C.T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinform.* **18**, 529 (2017).

55. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

56. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

57. Zheng, L. *et al.* Structural insight into the mechanism of energy transfer in cyanobacterial phycobilisomes. *Nat. Commun.* **12**, 5497 (2021).

58. Wilson, A. *et al.* Structural determinants underlying photoprotection in the photoactive orange carotenoid protein of cyanobacteria. *J. Biol. Chem.* **285**, 18364–18375 (2010).

Chapter III

Discussion

This thesis aimed at fathoming the paradox of molecular innovation and illustrated that functional novelty does not appear as paradoxical as it seems in the first place. In contrast, novel functional features evolved quite easily in the two investigated cyanobacteria light-perceiving model system.

The first original publication about the origin of innovative multi-color sensing in cyanobacteriochromes showed that the LCA of all CBCRs alternately sensed green and red incident light, in contrast to closely related canonical red and far-red sensing cyanobacterial phytochromes. The examined ancestral CBCRs (as the extant ones) function independently of adjacent domains with only a single GAF domain necessary for genuine light perception[1].

CBCRs are the only class of phytochromes that evolved the ability to sense the whole palette of the visible spectrum, whereas other members of the superfamily are limited to a narrow color range[2]. This indicates that unlocking the whole spectrum became possible only after uncoupling the function from adjacent domains in the polypeptide, meaning that the mutational space for tuning color perception in the CBCR GAF domains became accessible only after decoupling function from complex domain architecture. This represent a rather unusual example of a protein that evolved to execute a more sophisticated task with a less complex protein[3].

Multi-subunit proteins typically evolve into more complex structures over time, although this complexation is not mandatorily adaptive[3–5]. If this trend holds true for multi-domain proteins is arguable, but functional domain complexity evidently reversed in CBCRs. Our findings show that novel functional features like new color sensing do not necessitate more complex proteins, but can in fact be achieved via a simplification in protein domain structure.

However, we did not resolve the transition of a cyanobacterial knotless phytochrome into a CBCR on the substitutional level because the evolutionary distance between the two was too far to reliably do so. The evolutionary history of CBCRs is further characterized by countless gene duplications of entire polypeptides or single domains within the protein, horizontal transfers, domain fusions or fissions, and gene losses or conversions. Our evolutionary distance analysis between GAF domains on the same polypeptide showed that neighboring GAF domains can be evolutionary distant from each other and may even belong to different protein classes within the phytochromes, rendering it impossible to doubtlessly infer the evolutionary history of CBCR GAF domains with a single domain tree.

To account for these uncertainties, we inferred three possible LCA proteins of CBCR GAF domains that based on three plausible tree topologies with slight branch rearrangements near the root, and compared their characteristics. All reconstructed proteins agreed in the colors they sense and reinforced our hypothesis that green/red perception in the LCA of all CBCR GAF domains is a strong phylogenetic signal and independent from the exact branching order within the CBCR groups on the phylogeny.

However, due to these uncertainties in the exact branching order, we did not think that it was reasonable to work out which substitutions were necessary to change perceived colors in specific CBCR groups, but the short branch lengths between different groups on our phylogenies show that changing color perception may be achieved in only a few mutational steps in the single CBCR GAF domains. This shows that evolving novel color features could be mutationally quite simple in single domain phytochromes.

The ancestral CBCR proteins also reacted to blue-light irradiation, and were further able to bind an alternative chromophore (biliverdin) that is only used in some crown groups of extant CBCRs[6]. This blue-light sensing represents a typical moonlighting function[7]. It is imaginable that (if the corresponding gene got amplified under the right ecological conditions) mutations would have accumulated that could have turned the ancestral protein into a functional blue-light respondent CBCR. Additionally, the intrinsic latent affinity for biliverdin may also represent an exaptation for broader color tuning. Such evolutionary transition by exploiting latent moonlighting functions, would precisely follow the EAD model to evolve a novel functional feature[8]. If CBCRs really diversified like this, is still not clear. At present, we are limited by the phylogenetic uncertainties in CBCRs' history that prevent us from working out these evolutionary transitions on a mechanistic level of detail.

A future strategy to minimize phylogenetic uncertainties could be by reconciling the CBCR GAF domain history with individual histories of other phytochrome domains on the same polypeptide to potentially identify domain transfer, swapping, shuffling, or conversion events. Such reconciliation approach could result in a reliable single CBCR GAF domain tree that would facilitate finding specific color tuning residues. Together, this would help to understand the genetic basis of change in CBCR color perception, that we assume may be quite simple.

The second original publication investigated the evolution of a novel allosteric regulation in cyanobacterial photoprotection. This innovative feature of additional allosteric control was achieved by a foreign protein that was horizontally transferred into cyanobacteria from distantly related proteobacteria. This FRP-like protein was fully compatible with OCP1 the moment it got into cyanobacteria, because it exploited a conserved dimerization interface of OCP and the contacting residues of both proteins, FRP and OCP, were instantly and fortuitously fully compatible[9].

However, we cannot time the acquisition of FRP precisely, relative to the appearance of OCP1 paralogs in cyanobacteria. If FRP came in after OCP1 evolved, FRP would have been completely functional from the first encounter. If it came in before, FRP would have interacted with an ancestral OCP that existed at that time in a seemingly unproductive way: FRP binding to the ancestral OCP would have slowed down the photo-recovery reaction, consequently impeding fast recovery. However, this effect only appeared at molar excess of FRP, relative to the ancestral OCPs we tested.

It is not clear if such FRP to OCP ratios would have ever been reached inside a cyanobacterial cell. This means that the interaction would have either been neutral without selectable effect to the phenotype or, if the FRP to OCP ratios would have been reached and there was a physiological effect, it could otherwise explain the early and rapid diversification of OCP paralogs as an escape strategy to prevent slowing by FRP. Our AlphaFold2 predictions showed that extant OCPx paralogs could have indeed escaped from FRP binding in a single substitution from AncOCPall, presenting a plausible and simple one-step genetic trajectory to escape.

Another possible explanation would be an ancestral, less sophisticated mechanism of regulation: an OCP with bound FRP is not able to bind to the phycobilisome and consequently cannot perform its photoprotective task. This would represent a control mechanism that does not even necessitate a transformational change in OCP. Instead, OCP could always be present in the red, active form that would be inactivated by binding FRP in a dose-dependent manner. However, regulation of this kind of mechanism would only be possible on the transcriptional or translational level of FRP that would only be slow and further depend on additional gene or transcript regulation. Additionally, we showed that AncOCPall was already fully photo-switchable and did not need FRP for auto-regulation.

The immediately functional encounter of OCP and FRP in a cyanobacterial cell was totally accidental after a horizontal gene transfer event. If natural selection later helped to fix the *frp* gene in cyanobacteria, is unknown. To answer this, we would have to investigate the ecological conditions at that time, but possibilities to do so are limited[10]. The reason why the *frp* gene eventually was fixed in cyanobacteria will probably stay unanswered, as ecological information is not easily inferable phylogenetically.

The EAD model has recently been revised to include an additional step, called potentiation that accounts for the ecological opportunities to evolve novelty. Only if the genetic background conditions were favorable at the time, a gene coding for a protein with moonlighting function would be amplified and potentially fixed.[8] This further complicated PEAD model adds another condition to evolve novelty. In contrast, our study shows that the establishment of a novel allosteric control mechanism was primed by the introduction of an unrelated protein that was instantly compatible with the established cyanobacterial system just by chance. The horizontally transferred *frp* gene was fixed without further amplification or diversification, illustrating a much shorter route to establish biological novelty, that is simpler and eventually more parsimonious than the PEAD model.

The evolution of multi-color sensing in CBCRs could in turn perfectly be explained by the PEAD model, as marine cyanobacteria needed to perceive more colors to properly orient themselves in the water column to best use the incident sunlight for energy conversion while protecting from dangerously high irradiation. However, how the CBCR GAF domain became independent of adjacent domains that are necessary for genuine light perception in closely related phytochromes, is still unknown. This seemingly important step towards multi-color sensing could have also happened by chance, thus creating a functional green/red one-domain sensor fortuitously. Together with the evidence from the fortuitous encounter of FRP and OCP, we show that chance events may be an important first step in the evolution of functional novelty, before natural selection can even attack.

In most cases, it is impossible to tell what kind of evolutionary driving force or mechanism constructed a novel protein-protein interaction. Most proteins that we know to interact today originated from within the same proteome. This makes it impossible to tell apart if their compatibility was sudden and by chance or through a long genetic trajectory of increasing affinity, mainly driven by the power of natural selection.

In the OCP-FRP case, we could only tell that the surface compatibility evolved purely by chance, because the evidence for a horizontal gene transfer event was still inferable with molecular phylogenetics. This allowed to test the compatibility of OCP proteins with ancestral FRP proteins that evidently had existed before OCP and FRP have first met in an ancestral cyanobacterium. The compatible surface residues in both proteins, OCP and FRP, had evolved before in distantly related bacterial groups without a direct selection pressure for the future interaction. In OCP, the interaction surface is conserved since the LCA of all OCP and used to self-dimerize in the red state of the protein[11]. This ancient dimerization surface had then been molecularly exploited for the allosteric interaction by FRP that fortuitously came with the perfectly matching residues[12].

Although the new allosteric regulation was an innovative functional feature for cyanobacteria, no new protein had to be invented. Instead, evolution used already existing parts that were randomly mixed together and fortuitously fitting to create a functional novelty by chance. This is remindful of a tinkerer that invents new things by arranging already existing parts in a novel and productive way (perhaps sometimes also fortuitously)[13].

In analogy to the deep homology that explains the parallel evolution of vision and limbs[14], here, deep homology between proteins (FRPL and FRP) enabled the evolution of a novel functional feature by evolutionary tinkering even across species boundaries and through the priming power of blind chance. FRP's *de novo* function of allosteric regulation came as a hidden moonlighting function and may be considered a happy accident to cyanobacteria.

Finally, we must admit that we cannot identify the initiating evolutionary events in most molecular innovations. In the end, natural selection will probably determine which novelty to fix, and that is the result that we usually see or infer. But we have to be aware that priming chance events usually leave no traces and may thus be highly underrepresented in evolutionary studies, although chance may often be a simpler explanation for innovative novelty. We should consequently always consider the power of chance events in evolution, rather than taking the easy way out and categorically attributing biological innovation to purely adaptive processes.

References

1. Priyadarshini, N. *et al.* Evidence for an early green/red photocycle that precedes the diversification of GAF domain photoreceptor cyanobacteriochromes. *Photochemical & Photobiological Sciences* **22,** 1415–1427 (2023).

2. Rockwell, N.C. & Lagarias, J.C. A brief history of phytochromes. *ChemPhysChem.* **11,** 1172–1180 (2010).

3. Pillai, A.S., Hochberg, G.K.A. & Thornton, J.W. Simple mechanisms for the evolution of protein complexity. *Protein science* **31,** e4449 (2022).

4. Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F. & Gray, M.W. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63,** 528–537 (2011).

5. Hochberg, G.K.A. *et al.* A hydrophobic ratchet entrenches molecular complexes. *Nature* **588,** 503–508 (2020).

6. Fushimi, K. *et al.* Cyanobacteriochrome photoreceptors lacking the canonical Cys residue. *Biochemistry* **55,** 6981–6995 (2016).

7. Huberts, D.H.E.W. & van der Klei, I.J. Moonlighting proteins: an intriguing mode of multitasking. *Biochimica et biophysica acta* **1803,** 520–525 (2010).

8. Kassen, R. Experimental evolution of innovation and novelty. *Trends in Ecology & Evolution* **34,** 712–722 (2019).

9. Steube, N. *et al.* Fortuitously compatible protein surfaces primed allosteric control in cyanobacterial photoprotection. *Nature Ecology & Evolution* **7,** 756–767 (2023).

10. Schoonmaker, P.K. & Foster, D.R. Some implications of paleoecology for contemporary ecology. *Bot. Rev* **57,** 204–245 (1991).

11. Domínguez-Martín, M.A. *et al.* Structures of a phycobilisome in light-harvesting and photoprotected states. *Nature* **609,** 835–845 (2022).

12. Bridgham, J.T., Carroll, S.M. & Thornton, J.W. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312,** 97–101 (2006).

13. Jacob, F. Evolution and tinkering. *Science* **196,** 1161–1166 (1977).

14. Shubin, N., Tabin, C. & Carroll, S. Deep homology and the origins of evolutionary novelty. *Nature* **457,** 818–823 (2009).

Acknowledgements

In the first instance, I would like to thank Georg Hochberg for his excellent supervision and for offering the great opportunity to work on this fantastic topic of protein evolution.

I thank my thesis advisory committee (Georg Hochberg, Tobias Erb, Victor Sourjik) for fruitful discussions throughout my graduation and Peter Graumann, Thomas Friedrich, and Andreas Seubert for their willingness to participate in my thesis defense committee.

I further thank all my collaborators that made it possible to publish the two successful articles that represent the heart of this thesis.

Special thanks go to the Max Planck Foundation for giving me the opportunity to write most parts of this thesis at the very inspiring former summer house of Max Planck at lake Tegernsee.

All work in this thesis was kindly supported by the Max Planck Society.

The second article was moreover kindly supported by the European Research Council.

Finally, I especially thank all my former and present colleges in the Evolutionary Biochemistry research group for their great support from the very beginnings of the newly formed group until today.