A Case Study of Eukaryogenesis: The Evolution of Photoreception by Photolyase/Cryptochrome Proteins

Jennifer A. Miles\textsuperscript{1,2} · Thomas A. Davies\textsuperscript{1} · Robert D. Hayman\textsuperscript{1} · Georgia Lorenzen\textsuperscript{1} · Jamie Taylor\textsuperscript{1} · Mubeena Anjarwalla\textsuperscript{1} · Sammie J. R. Allen\textsuperscript{1} · John W. D. Graham\textsuperscript{1} · Paul C. Taylor\textsuperscript{1,2}

Received: 16 January 2020 / Accepted: 5 September 2020 / Published online: 26 September 2020

Abstract
Eukaryogenesis, the origin of the eukaryotes, is still poorly understood. Herein, we show how a detailed all-kingdom phylogenetic analysis overlaid with a map of key biochemical features can provide valuable clues. The photolyase/cryptochrome family of proteins are well known to repair DNA in response to potentially harmful effects of sunlight and to entrain circadian rhythms. Phylogenetic analysis of photolyase/cryptochrome protein sequences from a wide range of prokaryotes and eukaryotes points to a number of horizontal gene transfer events between ancestral bacteria and ancestral eukaryotes. Previous experimental research has characterised patterns of tryptophan residues in these proteins that are important for photoreception, specifically a tryptophan dyad, a canonical tryptophan triad, an alternative tryptophan triad, a tryptophan tetrad and an alternative tetrad. Our results suggest that the spread of the different triad and tetrad motifs across the kingdoms of life accompanied the putative horizontal gene transfers and is consistent with multiple bacterial contributions to eukaryogenesis.

Keywords Cryptochrome · Photolyase · Cyanobacteria · Horizontal gene transfer · Eukaryogenesis

Introduction
It is widely accepted that eukaryogenesis involved an endosymbiosis between at least one archaeon, probably from the Asgard lineage, and at least one bacterium, most likely an ancestral \( \alpha \)-proteobacterium (López-García and Moreira 2019). More controversial and less well understood is the possibility of further symbioses or other significant gene flows from bacteria to ancestral eukaryotes (Pittis and Gabaldón 2016; Spang et al. 2019). Both the order in which these key evolutionary events occurred and their timeline are unclear. Characterising the emergence of animals from this ancient mix is of particular interest and offers a useful perspective from which to explore the major transitions (Paps 2018).

In addition to the large-scale phylogenomic studies that underpin our emerging understanding of eukaryogenesis, the field may benefit from case studies of the evolution of individual genes/proteins across the kingdoms of life. Researchers are finding surprising similarities in chemical signalling-related proteins between animals and groups of complex bacteria, including cyanobacteria (Brash et al. 2014; Magnani et al. 2017; Miles et al. 2019; Millard et al. 2014; Picciano and Crane 2019; Ponting et al. 1999; Rawlings 2015) that pose some important questions about horizontal gene transfer (HGT) in the origin of animals. This offers an alternative perspective from which to view the major transitions in eukaryogenesis. A good object of study would therefore be a family of proteins found in many major kingdoms and taxa, including animals, and that possesses a highly distinctive biochemical motif or motifs that would allow us to further refine phylogenetic analyses.

The photolyase/cryptochrome superfamily is distinguished biochemically by the ability of these proteins to harbour radical-pair intermediates (Hore and Mouritsen 2016). This phenomenon allows photolyases and cryptochromes to
sense and react to sunlight, which is of enormous benefit to many species. There are two broad types of response, namely DNA repair and signalling (Mei and Dvornyk 2015). In particular, ultraviolet B (UVB) present in sunlight can be damaging to genetic material, changing the chemical structure of DNA, leading to aberrations such as cyclic pyrimidine dimers (CPDs) and (6–4) dimers. Once activated by light, photolyases are able to catalyse the reverse of some of these damaging reactions, restoring the original DNA structure (Sancar 2016). The chemical process is initiated by reduction of fully oxidised FAD to FAD− or FADH· and finally through to FADH−, with the reduced species providing electrons to cleave bonds in the defective DNA (Müller et al. 2015). Alongside this is an antenna chromophore with a different absorption range that broadens spectral properties. Five of these antenna chromophores have been identified so far for photolyase/cryptochrome superfamilies members (Kiontke et al. 2014). Alternatively, but based on the same chemistry, cryptochromes are core components of circadian rhythms, which allow organisms, including humans, to respond behaviourally to the daily cycle of light and dark, by initiating and shutting down biochemical processes as appropriate (Mei and Dvornyk 2015). Cryptochromes are also proposed to have a role in magnetoreception (Hore and Mouritsen 2016). Importantly for our work, the efficiency of the radical-pair mechanism appears to depend on precise details of the amino acid sequence of the cryptochrome in a given species, in particular a “dyad”, “triad” or “tetrad” of tryptophan residues (Müller et al. 2015; Cailliez et al. 2016). We aimed to carry out a wide-ranging phylogenetic analysis of the photolyase/cryptochrome superfamily, building on earlier work (see below), and to map the presence of tryptophan motifs across the resulting phylogenetic tree.

The aim of this study was to ascertain the bacterial contribution to the function and evolution of the photolyase/cryptochrome superfamily and hence inform models of eukaryogenesis. Previous phylogenetic analyses have taken place within this family (Kanai et al. 1997; Öztürk et al. 2007; Mei and Dvornyk 2015; Scheerer et al. 2015). The most recent comprehensive phylogenetic analysis of the photolyase/cryptochrome superfamily in eukaryotes was published by Mei and Dvornyk in 2015 (Mei and Dvornyk 2015), who noted the existence of prokaryotic orthologues but included only a few bacterial sequences in their analysis. Earlier, Lucas-Lledó and Lynch had undertaken another major phylogenetic analysis including numerous prokaryotic photolyase/cryptochromes (Lucas-Lledó and Lynch 2009). However, their study focused mainly on photolyase gene loss, whereas we are interested in the acquisition of particular biochemical features. Also, very many prokaryotic genomes have been sequenced since 2009. Therefore, we decided to start from the eukaryotic analysis from Mei and Dvornyk, to carry out our own search for prokaryotic orthologues and to cross-reference our findings to the work of Lucas-Lledó and Lynch. Previously it had been suggested that both gene duplication events, gene loss and convergent evolution may have taken place across this family of proteins, in particular during the evolution of cryptochromes from plants and animals, known as Plant Cry and Animal Cry, respectively (Cashmore et al. 1999).

Results

Phylogenetic Analysis of Currently Available Prokaryotic and Eukaryotic Genomes Supports the Existence of Seven Subfamilies of Photolyase/Cryptochromes

To start our search for candidate bacterial orthologues, we selected seven of the sequences used by Mei and Dvornyk (2015) as queries for their analysis. Listed in the Materials and Methods, these cover the previously identified subfamilies (6–4) Photolyase/Animal Cry, Plant Cry, Plant Photolyase, Cyclobutane Pyrimidine Dimer (CPD) Class I Photolyase, CPD Class II Photolyase and Cry-DASH (Fig. 1). The top 20 candidate orthologues from different bacterial species were selected from protein–protein BLAST (BLASTp) searches of the NCBI non-redundant protein database, filtered for taxid:2 (bacteria). We found no candidate bacterial orthologues of Plant Photolyase proteins using this method. In order to confirm that the identified sequences were candidate orthologues, these were then used in a reciprocal BLASTp search against the original species to see if the starting sequence was obtained as the best match. Sequences with complete genomes were kept and sequences obtained from metagenomes were removed when the genome completion was below 50%. Extra eukaryotic and archaeal sequences, beyond those used as start points for our BLASTp searches, were added to give a more complete picture of the photolyase/cryptochrome superfamily including species from most major taxa. We accrued 186 amino acid sequences in our database (full details in Supplementary Information).

According to eggNOG, 84% of orthologues of “cryptochrome/DNA photolyase, FAD-binding domain” are found in bacteria, 14% in eukaryota and 2% in archaea. The bacterial total of 84% comprises 45% from proteobacteria (15% α-, 6% β- and 22% γ-proteobacteria), 16% from actinobacteria, 8% from bacteroidetes/chlorobi, 5% from cyanobacteria, 5% from firmicutes and 5% other. A similar spread across the bacterial phyla is observed in OrthoDB. We have ensured there are representatives of all these phyla in our database.

We aligned our protein sequences using MUSCLE (Multiple Sequence Comparison by Log-Expectation) (Edgar 2004) to directly compare it to the analysis undertaken
by Mei and Dvornyk. To reinforce the reliability we also aligned our chosen sequences using MAFFT (Multiple Alignment using Fast Fourier Transform) (Katoh et al. 2017) and PRANK (Löytynoja and Goldman 2010) for further comparison. Following alignment and trimming of sequences using Gblocks (Castresana 2000), trees were generated using PhyML with approximate likelihood ratio test and Shimodaira-Hasegawa (aLRT-SH) used as an estimate for branch support in PhyML and the posterior probabilities shown for the Bayesian analysis. Again these are the same methods used by Mei and Dvornyk, with the exception that their alignment was trimmed manually using Bioedit, where we used Gblocks (Mei and Dvornyk 2015). The low stringency setting selected in Gblocks ensured that only isolated elements of particular sequences were removed. The LG (Le and Gascuel) model was used in all cases for tree generation, as chosen by smart model selection through the PhyML software (Le and Gascuel 2008).

The three resulting trees are presented in Figures S1 (MUSCLE), S2 (MAFFT) and S3 (PRANK). The previous analysis (Mei and Dvornyk 2015) aligned the sequences using MUSCLE. Hence we present the analysis based on MUSCLE here as our primary output. Our trees are presented unrooted, as is the case for trees from Lucas-Lledó and Lynch (2009) and Mei and Dvornyk (2015). The latter work devoted significant effort to establishing the geological timeline of the evolution of the photolyase/cryptochrome superfamily, giving confidence that CPD Class II Photolyases are basal to the superfamily. From our biochemical perspective, it is sensible to effectively use the CPD Class II Photolyases as an outgroup for the analysis, since, as discussed below, they only possess a “dyad” of tryptophan residues as compared with the triads and tetrads that are characteristic of the other subfamilies.

The (6–4) Photolyase/Animal Cry subfamily, the Cry-DASH subfamily and the Plant Photolyase subfamily appear in our Figures S1 and summary Fig. 2 much as they do in the equivalent tree from Mei and Dvornyk. The relative positions of the CPD Class I Photolyase and Plant Cry subfamilies show less consistency between our tree in Fig. 2 and Mei and Dvornyk’s results. However, the tree based on our MAFFT alignment (Fig. S2) exhibits more similarity. We should also note that the literature phylogenetic tree that includes more bacterial sequences (Lucas-Lledó and Lynch 2009) places the CPD Class I Photolyase and Plant Cry subfamilies quite differently. We conclude that our findings both broadly accord with the literature and reflect the unresolved subtleties of the previous analyses.

We should note an interesting set of orthologues from Chloroflexi bacteria. These five proteins form a monophyletic group in all three of our trees (Figs. S1–S3), but are located differently in each case with respect to the six subfamilies. When these sequences are modelled they appear to be most similar to CPD Class III Photolyases (these sequences align best with 4U63 from the CPD Class III Agrobacterium tumefaciens structure with 42% sequence identity between Roseiflexus orthologue and 43.38% homology with Chloroflexus aggregans orthologue) (Scheerer et al. 2015).

**Incongruences in the Evolution of the Superfamily**

The occurrence of horizontal gene transfer from ancestral bacteria to ancestral eukaryotes is widely accepted to have resulted from the endosymbioses that formed mitochondria and also, in the case of plants, chloroplasts. More recently, further examples of possible HGT have started to emerge. Indeed Lucas-Lledó and Lynch have mentioned the importance of gene gain, by HGT, as well as gene loss during the evolution of the photolyase/cryptochrome superfamily (Lucas-Lledó and Lynch 2009). However, these authors did not explore in detail how such HGT events contribute to the

---

**Fig. 1** The seven subfamilies used in phylogenetic analysis. Main representatives are shown alongside their predicted domains based on NCBI. PHR photolyase, FAD FAD-binding domain, CryC Cryptochrome C

| Subfamily          | PHR | FAD | CryC | Length |
|--------------------|-----|-----|------|--------|
| Animal Cry         |     |     |      |        |
| (6-4) Photolyase   | PHR |     |      | 586    |
| Cry-DASH           |     | FAD |      | 526    |
| CPD Class I        | PHR |     |      | 589    |
| CPD Class II       | PHR |     |      | 557    |
| Plant Cry          |     | FAD | CryC | 681    |
| Plant Photolyase   |     | FAD |      | 447    |
acquisition of new function during evolution and this is our primary focus.

To identify putative HGT occurrences, we have followed guidance from authors including Husnik and McCutcheon (2018) who conclude that "Phylogenetic conflict (that is, incongruence of a single-gene tree with a known species phylogeny) is the method of choice for the detection of HGT events." We have identified three significant incongruences in our photolyase/cryptochrome “single-gene tree” that are visible in all three Figures S1–S3. The relevant nodes are annotated in summary Fig. 2 as N1, N2 and N3. Node 1 (N1) represents a common ancestor of photolyases and cryptochromes from various taxa of bacteria, from archaea, from algae, fungi, plants and animals. However, this appears more recently than the common ancestor of animal CPD Class II Photolyase proteins and the other animal paralogues in the superfamily. This is clearly incongruent with the standard tree of life. Node 2 (N2) represents the most recent common ancestor of Cry-DASH proteins from a group of eukaryotes, including animals, and of bacteria, including numerous cyanobacterial proteins. Again, the standard model would not predict this to fall more recently than the common ancestor of the eukaryotic members of the superfamily. Node 3 (N3) represents the most recent common ancestor of cryptochromes (Cry) from animals, fungi, plants, algae and, remarkably, two species of Gloeobacter, a further striking incongruence. This last feature was noted by Lucas-Lledó and Lynch (2009), though G. violaceus was the only Gloeobacter genome then available. They invoked a HGT from green algae to G. violaceus.

The trees from Lucas-Lledó and Lynch (2009) and Mei and Dvornyk (2015) show a node, similar to our Node 1, that represents a common ancestor of all cryptochrome/photolyase proteins, except CPD Class II Photolyases, in most major phyla of bacteria and eukaryotes. They also both show a node, similar to our Node 2, that represents a common ancestor of (6–4) Photolyase/Animal Cry and bacterial Cry-DASH sequences (though Lucas-Lledó and Lynch did not label this group as Cry-DASH). Our tree is the first to include sequences from both sequenced species of Gloeobacter, but the location of Node 3 accords with the placement of G. violaceus in the literature trees. These previous authors did not discuss these apparent incongruences in detail.

We should also note the presence of archaeal sequences in our data set. Apart from one sequence from Thaumarchaeota, all the archaeal species represented are Euryarchaeota, in which HGT from bacteria is a common feature (see for example Santa-Molina et al. 2020). Therefore, the presence of photolyase/cryptochrome superfamily genes in these archaea is not informative with respect to the analysis presented here.

Conserved Residues and Hence Functionalities are Observed Across All Families

Previously Zhang et al (2017) identified conserved residues in the superfamily that relate to DNA binding within the photolyase, defining which substrate is recognised and repaired by the photolyase (Table 1 and Fig. 3). In Fig. 3 they are labelled as position 1 and position 2. Looking at position 1 we can see that, like their eukaryotic counterparts, bacterial orthologues of CPD Photolyase proteins have a
methionine in this position. The cyanobacterial *Gloeobacter* sequences that surprisingly group with (6–4) Photolyase/Animal Cry have the characteristic histidine at the same position and bacterial orthologues of Cry-DASH have the expected glutamine.

Conversely, in position 2, all of the members of the CPD Photolyase subfamilies have a tryptophan at this position apart from the fungal CPD Photolyase, which, in comparison to its bacterial orthologue, contains a tyrosine. The bacterial orthologues of Cry-DASH retain the tyrosine also at position 2. The *Gloeobacter* sequences again have the conserved tryptophan seen in the rest of the (6–4) Photolyase/Animal Cry subfamily. The bacterium *Xanthomonas axonopodis* that groups with the Plant Cry subfamily appears to have residues more like a CPD Photolyase at these positions (M358 and W397). The Plant Photolyase member doesn’t convey any conserved residues at these positions. Conservation of these residues in the bacterial sequences in our study gives confidence that the orthologues are genuine and that they will have the same function in vivo as eukaryotic members of the same subfamily.

The “Canonical” Tryptophan Triad and Tryptophan Tetrad Motifs have Emerged Since the Divergence with CPD Class II Photolyases

Crucial to the function of some members of this superfamily is a triad of tryptophan residues that are required for the transmission of the electron from FAD to create FADH which is then, for example, used for the repair of DNA in (6–4) Photolyases (Hitomi et al. 2009). We mapped the presence of these three tryptophans on our phylogenetic tree and observe that this mechanism appears to have evolved more recently than the divergence from CPD Class II Photolyase proteins (Fig. 4). The three tryptophan residues are present in the majority of orthologues that have evolved since Node 1. In subsequent research, the tryptophan triad that is involved in electron transfer to FAD has been expanded to include a further tryptophan in *Drosophila* and *Xenopus* (Müller et al. 2015; Nohr et al. 2016; Yamamoto et al. 2017; Lin et al. 2018). This tryptophan “tetrad” increases the lifetime of FADH. The fourth tryptophan in the tetrad is replaced by a tyrosine in algae and plants, which is able to produce a similar effect (Nohr et al. 2016; Oldemeyer et al. 2016; Franz et al. 2018).

We wondered when this tryptophan tetrad motif evolved, so looked for the presence of the fourth tryptophan throughout our phylogenetic analysis. The sequences that contain this fourth tryptophan are shown on the phylogenetic tree in Fig. 4 with red branches. All of the animal cryptochromes have this fourth tryptophan, usually in a four amino acid motif of W(Φ)SW, where the first Trp is the newly identified member of the “tetrad”. We also identified this motif in the two sequenced *Gloeobacter*, which are cyanobacteria (*Gloeobacter kilaeuensis* WP_023172734.1 and *Gloeobacter violaceus* WP_011141747.1) (Figs. 5 and 6). These proteins group near the algal (6–4) Photolyases, which are

| Subfamily   | Position 1 | Position 2 | Substrate |
|-------------|------------|------------|-----------|
| (6–4) photolyase | H          | W          | 6–4       |
| Animal Cry   | H          | W          | –         |
| CPD class I  | M          | W          | CPD       |
| CPD class II | M          | W          | CPD       |
| CPD class III| M          | W          | CPD       |
| Cry-DASH     | Q          | Y          | ssDNA     |
| Plant Cry    | V          | Y          | –         |
| Plant photolyase | –            | –            |           |

### Table 1 Conserved residues identified across members of the superfamily

- **Xenopus laevis CPD**
- **Geobacter sulfurreducens CPD**
- **Agrobacterium tumefaciens CPD**
- **Arabidopsis thaliana Cry1**
- **Xanthomonas axonopodis**
- **Arabidopsis thaliana PHR2**
- **Verticillium alboae Valles, 1902 CPD**
- **Cronobacter sakazakii CPD**
- **Homo sapiens Cry1**
- **Xenopus laevis SML**
- **Gloeobacter kilaeuensis PL**
- **Xenopus laevis CD**
- **Synechocystis PCC 6803 CD**

**Fig. 3** Alignment of examples of each subfamily alongside bacterial orthologues for each (aside from CPD III which is characterised from the bacteria *A. tumefaciens* and from Plant Photolyase *A. thaliana* which has no bacterial orthologues). Residues involved in substrate recognition are highlighted with a black box and whole conserved residues are shown with an *. Sequences used are *X. laevis* (6–4) PL (NP_001081421.1), *X. laevis* CPD II (NP_001089127.1), *X. laevis* CD (NP_001084438.1), *Geobacter sulfurreducens* CPD II (WP_010943456.1), *Agrobacterium tumefaciens* (WP_010971478.1), *A. thaliana* Cry1 (NP_567341.1), *A. thaliana* PHR2 (NP_182281.1), *X. axonopodis* (WP_011050942.1), *V. alboae* (XP_002999933.1), *Cronobacter sakazakii* WP_012125345.1, *H. sapiens* Cry1 (NP_004066.1), *G. kilaeuensis* (WP_023172734.1), *Synechocystis* PCC 6803 CD (WP_014407097.1) (Color figure online)
able to undertake DNA repair and gene control using an unusual arrangement of long C-terminal extension (Kottke et al. 2017; Franz et al. 2018). As mentioned above, this provides a firm biochemical justification for treating CPD Class II Photolyases as, essentially, an outgroup in our phylogenetic analysis.

An “Alternative” Tryptophan Triad is Characteristic of Cry-DASH Proteins

Whilst the canonical triad (and tetrad) of tryptophan residues is characteristic of all subfamilies of the photolyase/cryptochrome superfamily except CPD Class II Photolyases, experiments in *Escherichia coli* and *Arabidopsis thaliana* have shown that this particular triad is not always essential for function (Y. F. Li et al. 1991; X. Li et al. 2011). Indeed, Biskup et al. (2011) showed in the cyanobacterium *Synechocystis* that an “alternative” triad of tryptophan residues could mediate electron transfer (Fig. 6b). Two of the canonical tryptophan residues still contribute to the alternative set, but in *Synechocystis* a W320 completes the triad, not W375. In some orthologues, the redundant tryptophan is lost.

In Figures S5 and S7, sequences containing the “alternative” triad of tryptophan residues are indicated with green coloured branches (Fig. 7). Importantly, the identification of the different triads in our work rests solely on the alignment of the sequences. In many cases, the proteins possess the elements of both the canonical and alternative triads and, with the exception of *Synechocystis* (see above), we do not know which of them is functional. The alternative triad is found sporadically amongst CPD Class I Photolyases, including those that use a different cofactor (5,10-methylenetetrahydrofolate, MTHF) and that are often called CPD Class III Photolyases (Scheerer et al. 2015). The alternative pattern is not found at all in Plant Photolyase proteins, nor in (6–4) Photolyase/Animal Cry.

Strikingly, the alternative tryptophan triad appears to be an essential feature of Cry-DASH proteins, being present in all the Cry-DASH sequences in our phylogenetic tree, with the exception of the proteins from diatoms *Thalassiosira pseudodonna* and *Phaeodactylum tricornutum* (Figure S5). In fact, the *P. tricornutum* protein has been shown experimentally to more closely resemble a Plant Photolyase in function, hence the annotation as Cry-DASH could be revisited (Juhás et al. 2014).

**Discussion**

We present here a phylogenetic analysis of cryptochrome and photolyase proteins that includes an array of prokaryotic sequences. The tree in Fig. 2 has all the hallmarks of the previous definitive analysis of this superfamily from Mei
and Dvornyk. However, similar to our previous research on retinoic acid biosynthesis (Miles et al. 2019; Millard et al. 2014), we find that inclusion of photolyase/cryptochrome sequences and candidate orthologues from across the kingdoms of life brings to the fore phylogenetic relationships that are particularly relevant to eukaryogenesis, in particular with respect to the placement of animal and bacterial orthologues.

We have highlighted three striking incongruences between our gene tree for the photolyase/cryptochrome superfamily and the widely accepted species tree that is based on analysis of ribosomal protein sequences (Hug et al. 2016). The simplest hypothesis that accords with this observation is to propose at least three HGT events between ancestral bacteria and ancestors of eukaryotes that have contributed to the emergence of this superfamily of proteins.

Mapping the occurrence of tryptophan dyads, canonical and alternative triads and tetrads across the phylogenetic tree is revealing. Broadly, we can say that CPD class II photolyase proteins possess a dyad and are thus predicted to be the least effective at stabilising diradicals. Indeed, we are essentially using the CPD Class II Photolyase sequences as an outgroup in this study. We then see acquisition of triad motifs that approximately coincides with the putative HGT at Node 1. Strikingly, within the Cry-DASH subfamily that has a common ancestor at Node 2, we find that the “alternative” triad is completely conserved, again indicating this feature may have accompanied HGT. Finally, around Node 3, where we propose a further HGT, we see appearance of tryptophan tetrad motifs. These tryptophan signatures provide powerful additional support for multiple HGT events.
In Fig. 2, we highlight where candidate bacterial orthologues from two key groups of bacteria, namely α-proteobacteria and cyanobacteria, are located, to clarify how the incidences of HGT we propose relate to the widely accepted endosymbiotic gene transfer (EGT) events, that is EGT between an ancestral α-proteobacterium and an ancestor of all eukaryotes and EGT between an ancestral cyanobacterium and an ancestor of plants. From an animal perspective, if we ignore isolated branches for sequences from each of *Tetrahymena thermophila* and *Trypanosoma brucei*, we find animal photolyase/cryptochrome superfamily proteins in only three clades, namely CPD Class II Photolyase, Cry-DASH and (6–4) Photolyase/Animal Cry. If EGT is the source of bacterial genes in animals, then we would expect to observe animal sequences having more recent common ancestors with α-proteobacterial proteins than with other bacterial groups in these CPD Class II Photolyase, Cry-DASH and (6–4) Photolyase/Animal Cry clades.

In fact, in our tree, α-proteobacterial sequences are only found in the Plant Cry clade. Lucas-Lledó and Lynch did identify some α-proteobacterial orthologues of Cry-DASH in their study, but they were not recovered by our methods. This is easy to understand from a simple BLASTp of Cry-DASH from, for example, *X. laevis*. There are hundreds of hits from cyanobacteria, firmicutes, CFB group bacteria etc. that score more highly than the closest α-proteobacterial match.

So, if we focus on animals, α-proteobacteria and cyanobacteria, we find that animal (6–4) Photolyase/Animal Cry genes have a more recent common ancestor with cyanobacteria (specifically two *Gloeobacter* species) than with α-proteobacteria (see Node 3) and we find that animal Cry-DASH genes also have a more recent common ancestor with cyanobacteria than with α-proteobacteria (see Node 2). Gabaldón (2018) and López-García and Moreira (2020) argue that is very difficult to reconcile such observations with just an EGT from an α-proteobacterium to form the mitochondrion. We would need to assume that orthologues of Animal Cry and of animal Cry-DASH genes had been lost in all sequenced extant species of α-proteobacterium (with the exception of those mentioned above where there are “distant” orthologues). Rather, evidence is accumulating for “waves” of acquisition of genes from different bacterial sources in addition to the endosymbiosis to form the mitochondrion and Gabaldón (2018) says “we need to go beyond simple models that involve a single bacterial endosymbiont engulfed by an archaeal ancestor”.

Intriguingly, the two *Gloeobacter* species in our analysis have photolyase/cryptochrome proteins that exhibit all of the canonical triad, the alternative triad and the tetrad of tryptophan residues (Fig. 8). The *Gloeobacter* proteins with the tryptophan tetrad are annotated as (6–4) Photolyases in both *G. kilaeuensis* and *G. violaceus* and the proteins with the characteristic alternative triad are annotated, consistent with our observation on the Cry-DASH group above, as Cry-DASH in both species. A further pair of *Gloeobacter* proteins group with a number of CPD Class I Photolyases that are believed to use 8-HDF as the second cofactor. Finally, *G. kilaeuensis* has yet another, unannotated, paralogue.

*Gloeobacter* species are generally accepted to have branched from the rest of the cyanobacteria very early in evolutionary history (Schirrmeister et al. 2013). *Gloeobacter* species lack thylakoid membranes that are the normal sites in cyanobacteria of light-dependent reactions. *G. kilaeuensis* was discovered in a Hawaiian lava cave in 2013 (Saw et al. 2013), whilst *G. violaceus PCC 7421* was isolated from a

---

**Fig. 6** a Model of (6–4) Photolyase orthologue from *G. kilaeuensis* using Phyre 2. The conserved tryptophan tetrad is shown in blue with FAD modelled in from 3FY4 shown in peach. b Model of a Cry-DASH orthologue from *G. kilaeuensis* with the “alternative” tryptophan triad shown in blue (Color figure online)
limestone rock surface (Nakamura et al. 2003; Mareš et al. 2013). To confirm that the seven photolyase/cryptochrome genes most probably evolved as paralogues in *Gloeobacter*, the GC contents of the genes were compared with that of the overall organism. All seven of the photolyase/cryptochrome genes in *Gloeobacter* have GC contents that are consistent with that of the whole organism (Table 2), which is evidence against acquisition of these genes by a more recent HGT to the cyanobacteria, as proposed by Lucas-Lledó and Lynch (2009).

Overall, our analysis suggests that bacteria have contributed significantly and repeatedly to the evolution of photolyases and cryptochromes across all kingdoms of life, including in animals, by putative HGT. These apparent gene transfer events seem to be linked to the acquisition of additional and/or different tryptophan residues that contribute to increased stability of the diradical intermediate that is so crucial to the mechanism of photoreception. We are mindful of the dangers of inferring features of species trees from single gene trees (Ku et al. 2015) and that is not our intention. Rather we aim to show that incongruent nodes in this gene tree align with fundamental distinctions in the biochemistry of different groups of photolyase/cryptochrome proteins, which invites further investigation. Indeed, case studies such as the complex phylogeny of the photolyase/cryptochrome superfamily could contribute valuable insight to emerging models of eukaryogenesis that acknowledge a rich bacterial heritage.

**Materials and Methods**

**Phylogenetic Analyses**

Cryptochrome and photolyase starting sequences were chosen from (Mei and Dvornyk 2015) and BLASTp searched against the NCBI non-redundant protein database against bacteria with the default settings to identify orthologous proteins. The initial sequences were Animal CRY (*Homo sapiens* NP_004066 and NP_066940), (6–4) Photolyase (*X. laevis* NP_001081421), CPD Class II Photolyase (*Monosiga brevicollis* MX1 XP_001746666), Cry-DASH (*Salpingoeca* sp. ATCC 50,818 XP_004989008), Fungal CPD Class I Photolyase (*Verticillium alfalfae* VaMs.102 XP_002999933); Plant Cry (*A. thaliana* NP_567341 and NP_171935); Plant Photolyase (*A. thaliana* NP_182281). The top 20 bacterial orthologues for all subfamilies were chosen, along with orthologues from all other groups of life. Full details of sequences collected are in the Supplementary Information. Some sequences were removed when they were found to be from incomplete genome sequences such as those derived from metagenomes.

Amino acid sequences were aligned using MAFFT with method FFT-NS-I (Katoh et al. 2017), MUSCLE (Edgar 2004) and PRANK (Löytynoja and Goldman 2010). Redundant regions of the alignment were then removed using Gblocks with low stringency settings (Minimum number of seq for a conserved and flank position set to half, Maximum number of contiguous nonconserved positions: 50, Minimum length of a block: 2, All gap positions allowed) (Castresana 2000). This retained 739 positions out of 2049 with the MUSCLE alignment (36% in 5 blocks), 732 positions out of 2075 with the MAFFT alignment (38% in 6 blocks) and 697 positions out of 5842 in the PRANK alignment (12% in 10 blocks). Phylogenetic trees were then built using PhyML
with approximate likelihood ratio testing giving branch support values and the Le and Gascuel (LG) model used which had been chosen in Smart Model Selection (Le and Gascuel 2008; Guindon et al. 2010; Lefort et al. 2017). Trees were analysed and annotated using TreeGraph2 (Stöver and Müller 2010).

**GC Content Calculation**

Genomes in the NCBI were used to identify the GC content of the entire genome of a bacteria and compare this to the GC content of the gene. A discrepancy of 8% was the cutoff set for a gene to be ‘real’ and not a possible contaminant.

**Modelling Cyanobacterial Proteins**

The structure of *G. kilaueensis* (WP_023172734.1) and *G. violaceus* (WP_011141747.1) (6–4) PL were modelled using Phyre 2.0 (Kelley et al. 2015) with 3FY4 as the template (Hitomi et al. 2009). The Cry-DASH orthologues were modelled with 1NP7 as the template (*G. kilaueensis* WP_023173156.1 and *G. violaceus* WP_011140837.1) (Brudler et al. 2003). The 8-HDF like orthologues were modelled with 1TEZ as the template (*G. kilaueensis* WP_023172026.1 and *G. violaceus* WP_011140786.1) (Mees et al. 2004). Finally, the other *G. kilaueensis* orthologue was modelled with 4U63 as a template (WP_023175030.1) (Scheerer et al. 2015).

**Acknowledgements**

This work was undertaken on MARC1, part of the High Performance Computing and Leeds Institute for Data Analytics (LIDA) facilities at the University of Leeds, UK. Martin Callaghan from the HPC at University of Leeds for help with establishing methods. We also thank Professor Mary O’Connell from the University of Nottingham for her input during the development of this project.

**Funding**

Open access funding provided by University of Leeds.

**Data Availability**

Data provided as Supplementary Information.
Compliance with Ethical Standards

Conflict of interest  Paul Taylor is a Director of Tangent Reprofiling Limited. Other authors have no competing financial interests.

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

References

Biskup T, Hitomi K, Getzoff ED et al (2011) Unexpected electron transfer in cryptochrome identified by time-resolved EPR spectroscopy. Angew Chemie - Int Ed. https://doi.org/10.1002/anie.201104321

Brash AR, Niraula NP, Boeglin WE, Mashhadi Z (2014) An ancient relative of cyclooxygenase in cyanobacteria is a linolate 10S-dioxygenase that works in tandem with a catalase-related protein with specific 10S-hydroperoxide lyase activity. J Biol Chem. https://doi.org/10.1074/jbc.M114.555904

Brudler R, Hitomi K, Daiyasu H et al (2003) Identification of a new cryptochrome class. Mol Cell. https://doi.org/10.1016/S1097-2765(03)00008-X

Cailliez F, Müller P, Firmino T et al (2016) Energetics of photodynamic charge migration within the tryptophan tetrad of an animal (6–4) photolyase. J Am Chem Soc. https://doi.org/10.1021/jacs.5b10938

Cashmore AR, Jarillo JA, Wu YJ, Liu D (1999) Cryptochromes: blue light receptors for plants and animals. Science. https://doi.org/10.1126/science.284.5415.760

Castrasana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh340

Franz S, Ignatz E, Wenzel S et al (2018) Structure of the bifunctional cryptochrome aCRY from Chlamydomonas reinhardtii. Nucleic Acids Res. https://doi.org/10.1093/nar/gky621

Gabaldón T (2018) Relative timing of mitochondrial endosymbiosis and the “pre-mitochondrial symbioses” hypothesis. IUBMB Life. https://doi.org/10.1002/ubl.1950

Guindon S, Duфayard JF, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. https://doi.org/10.1093/sysbio/syq010

Hitomi K, DiТаччio L, Arvai AS et al (2009) Functional motifs in the (6–4) photolyase crystal structure make a comparative framework for DNA repair photolyases and clock cryptochromes. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.0809180106

Hore PJ, Mouritsen H (2016) The radical-pair mechanism of magnetoreception. Annu Rev Biophys. https://doi.org/10.1146/annurev-biophys-032116-094545

Hug LA, Baker BJ, Anantharaman K et al (2016) A new view of the tree of life. Nat Microbiol. https://doi.org/10.1038/nmicrobiol.2016.48

Husnik F, McCutcheon JP (2018) Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. https://doi.org/10.1038/nrmicro.2017.137

Juhas M, Von Zadow A, Spexard M et al (2014) A novel cryptochrome in the diatom Phaeodactylum tricornutum influences the regulation of light-harvesting protein levels. FEBS J. https://doi.org/10.1111/febs.12782

Kanai S, Kikuno R, Toh H et al (1997) Molecular evolution of the photolyase-blue-light photoreceptor family. J Mol Evol. https://doi.org/10.1007/pl00006258

Katoh K, Rozewicki J, Yamada KD (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. https://doi.org/10.1093/bib/bbx108

Kelley LA, Mezulis S, Yates CM et al (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. https://doi.org/10.1038/nprot.2015.053

Kiontke S, Geiselbrecht Y, Pokorny R et al (2011) Crystal structures of an archaeal class II DNA photolyase and its complex with UV-damaged duplex DNA. EMBO J. https://doi.org/10.1038/emboj.2011.313

Kiontke S, Gnaу P, Haselsberger R et al (2014) Structural and evolutionary aspects of antenna chromophore usage by class II photolyases. J Biol Chem. https://doi.org/10.1074/jbc.m113.542431

Kottke T, Oldemeyer S, Wenzel S et al (2017) Cryptochrome photoreceptors in green algae: unexpected versatility of mechanisms and functions. J Plant Physiol. https://doi.org/10.1016/j.jplph.2017.05.021

Ku C, Nelson-Sathi S, Roettger M et al (2015) Endosymbiotic origin and differential loss of eukaryotic genes. Nature. https://doi.org/10.1038/nature14963

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol. https://doi.org/10.1093/molbev/msn067

Lefort V, Longueville JE, Gascuel O (2017) SMS: smart model selection in PhyML. Mol Biol Evol. https://doi.org/10.1093/molbev/msx149

Li YF, Heelis PF, Sancar A (1991) Active site of DNA photolyase: tryptophan-306 is the intrinsic hydrogen atom donor essential for flavin radical photoreduction and DNA repair in vitro. Biochemistry. https://doi.org/10.1021/bi00239a034

Li X, Wang Q, Yu X et al (2011) Arabidopsis cryptochrome 2 (CRY2) functions by the photoactivation mechanism distinct from the tryptophan (tp) triad-dependent photoreaction. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1114579108

Lin C, Top D, Manahan CC et al (2018) Circadian clock activity of cryptochrome relies on tryptophan-mediated photoreduction. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1702293114

López-García P, Moreira D (2020) The Syntrophy hypothesis for the origin of eukaryotes revisited. Nat Microbiol. https://doi.org/10.1038/s41564-019-0495-5

López-García P, Moreira D (2020) The Syntrophy hypothesis for the origin of eukaryotes revisited. Nat Microbiol. https://doi.org/10.1038/s41564-020-0710-4

Löytynoja A, Goldman N (2010) WebPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-11-579

Lucas-Lledó JI, Lynch M (2009) Evolution of mutation rates: phylogenetic aspects of antenna chromophore usage by class II photolyases. J Mol Evol. https://doi.org/10.1007/s00239-009-9276-2

Magnani F, Nenci S, Millana Fananas E et al (2017) Crystal structures of an archaeal class II DNA photolyase and its complex with UV-damaged duplex DNA. EMBO J. https://doi.org/10.1038/embr.2011.313

Mareš J, Hrouzek P, Kaňa R et al (2013) The primitive thylakoid-lacking cyanobacterium Gloeobacter is a common rock-dwelling
organism. PLoS ONE. https://doi.org/10.1371/journal.pone.00663
23
Mees A, Klar T, Gnaau P et al (2004) Crystal structure of a photolyase
bound to a CPD-like DNA lesion after in situ repair. Science. https :
://doi.org/10.1126/science.1101598
Mei Q, Dvornyk V (2015) Evolutionary history of the photolyase/
cryptochrome superfamily in eukaryotes. PLoS ONE. https://doi.
https://doi.org/10.1371/journal.pone.0135940
Miles JA, Machattou P, Nevin-Jones D et al (2019) Identification of
a cyanobacterial aldehyde dehydrogenase that produces reti-
noic acid in vitro. Biochem Biophys Res Commun. https://doi.
://doi.org/10.1016/j.bbrc.2018.12.171
Millard A, Scanlan DJ, Gallagher C et al (2014) Unexpected evolution-
ary proximity of eukaryotic and cyanobacterial enzymes responsi-
able for biosynthesis of retinoic acid and its oxidation. Mol Biosyst.
https://doi.org/10.1039/c3mb70447e
Müller P, Yamamoto J, Martin R et al (2015) Discovery and func-
tional analysis of a 4th electron-transferring tryptophan conserved
exclusively in animal cryptochromes and (6–4) photolyases. Chem
Commun. https://doi.org/10.1039/c5cc06276d
Nakamura Y, Kaneko T, Sato S et al (2003) Complete genome structure
of Gloeobacter violaceus PCC 7421, a cyanobacterium that lacks
thylakoids (supplement). DNA Res. https://doi.org/10.1093/dnare
s/10.4.137
Nohr D, Franz S, Rodriguez R et al (2016) Extended electron-transfer
in animal cryptochromes mediated by a tetrad of aromatic amino
acids. Biophys J. https://doi.org/10.1016/j.bpj.2016.06.009
Oldemeyer S, Franz S, Wenzel S et al (2016) Essential role of an
unusually long-lived tyrosyl radical in the response to red light
of the animal-like cryptochrome acry. J Biol Chem. https://doi.
://doi.org/10.1074/jbc.M116.726976
Öztürk N, Song SH, Özgür S et al (2007) Structure and function of ani-
mal cryptochromes. Cold Spring Harbor Symposia on quantitative
biology. Cold Spring Harbor Laboratory Press, New York. https :
://doi.org/10.1101/sqhb.2007.72.015
Paps J (2018) What makes an animal? The molecular quest for the ori-
gin of the Animal Kingdom. Integr Comp Biol. https://doi.
org/10.1093/icb/icx036
Picciano AL, Crane BR (2019) A nitric oxide synthase-like protein
from Synechococcus produces NO/NO$_3^-$ from L-arginine and
NAPDH in a tetrahydrobiopterin- and Ca$^{2+}$-dependent manner. J
Biol Chem. https://doi.org/10.1074/jbc.RA119.008399
Pitts AA, Gabaldón T (2016) Late acquisition of mitochondria by
a host with chimaeric prokaryotic ancestry. Nature. https://doi.
://doi.org/10.1038/nature16941
Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV (1999) Eukar-
yotic signalling domain homologues in archaea and bacteria.
Ancient ancestry and horizontal gene transfer. J Mol Biol. https :
://doi.org/10.1006/jmbi.1999.2827
Rawlings ND (2015) Bacterial calpains and the evolution of the calpain
(C2) family of peptidases. Biol Direct. https://doi.org/10.1186/
s13062-015-0095-0
Santa-Molina C, Rivas-Marín E, Rojas AM et al (2020) Origin and
evolution of polycyclic triterpene synthesis. Mol Biol Evol. https :
://doi.org/10.1093/molbev/msaa054
Saw IHW, Schatz M, Brown MV et al (2013) Cultivation and complete
gene sequencing of Gloeobacter kilaueensis sp. Nov., from a
Lava Cave in Kilauea Caldera, Hawai‘i. PLoS ONE. https://doi.
://doi.org/10.1371/journal.pone.0076376
Sancar A (2016) Mechanisms of DNA repair by photolyase and exci-
sion nuclease (nobel lecture). Angew Chem - Int Ed. https://doi.
://doi.org/10.1002/anie.201601524
Scheerer P, Zhang F, Kalms J et al (2015) The class III cyclobutane
pyrimidine dimer photolyase structure reveals a new antenna
chromophore binding site and alternative photoreduction path-
ways. J Biol Chem. https://doi.org/10.1074/jbc.M115.637868
Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC (2013) Evolu-
tion of multicellularity coincided with increased diversification
of cyanobacteria and the great oxidation event. Proc Natl Acad Sci.
https://doi.org/10.1073/pnas.1209927110
Spang A, Stairs CW, Dombrowski N et al (2019) Proposal of the
reverse flow model for the origin of the eukaryotic cell based on
comparative analyses of Asgard archaeal metabolism. Nat Micro-
biol. https://doi.org/10.1038/s41564-019-0406-9
Stöver BC, Müller KF (2010) TreeGraph 2: combining and visualizing
evidence from different phylogenetic analyses. BMC Bioinform.
https://doi.org/10.1186/1471-2105-11-7
Yamamoto J, Shimizu K, Kanda T et al (2017) Loss of fourth elec-
tron-transferring tryptophan in animal (6–4) photolyase impairs
DNA repair activity in bacterial cells. Biochemistry. https://doi.
://doi.org/10.1021/acs.biochem.7b0366
Zhang M, Wang L, Zhong D (2017) Photolyase: dynamics and elec-
tron-transfer mechanisms of DNA repair. Arch Biochem Biophys.
https://doi.org/10.1016/j.abb.2017.08.007