ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Photoglobin, a distinct family of non-heme binding globins, defines a potential photosensor in prokaryotic signal transduction systems

Theresa Schneider [a], Yongjun Tan [a], Huan Li [a], Jonathan S. Fisher [a], Dapeng Zhang [a,b,*]

[a] Department of Biology, College of Arts & Sciences, Saint Louis University, Saint Louis, MO 63105, United States
[b] Program of Bioinformatics and Computational Biology, College of Arts & Sciences, Saint Louis University, MO 63103, United States

ARTICLE INFO

Article history:
Received 28 August 2021
Received in revised form 14 December 2021
Accepted 15 December 2021
Available online 18 December 2021

Keywords:
Globin
Evolution
Phycocyanin
Photoreceptor
Signal transduction
B12-binding

ABSTRACT

Globins constitute an ancient superfamily of proteins, exhibiting enormous structural and functional diversity, as demonstrated by many heme-binding families and two non-heme binding families that were discovered in bacterial stressosome component RsbR and in light-harvesting phycobiliproteins (phycocyanin) in cyanobacteria and red algae. By comprehensively exploring the globin repertoire using sensitive computational analyses of sequences, structures, and genomes, we present the identification of the third family of non-heme binding globins—the photoglobin. By conducting profile-based comparisons, clustering analyses, and structural modeling, we demonstrate that photoglobin is related to, but distinct from, the phycocyanin family. Photoglobin preserves a potential ligand-binding pocket, whose residue configuration closely resembles that of phycocyanin, indicating that photoglobin potentially binds to a comparable linear tetrapyrrole. By exploring the contextual information provided by the photoglobin's domain architectures and gene-neighborhoods, we found that photoglobin is frequently associated with the B12-binding light sensor domain and many domains typical of prokaryotic signal transduction systems. Structural modeling using AlphaFold2 demonstrated that photoglobin and B12-binding domains form a structurally conserved hub among different domain architecture contexts. Based on these strong associations, we predict that the coupled photoglobin and B12-binding domains act as a light-sensing regulatory bundle, with each domain sensing different wavelengths of light resulting in switch-like regulation of downstream signaling effectors. Thus, based on the above lines of evidence, we present a distinct non-heme binding globin family and propose that it may define a new type of light sensor, by means of a linear tetrapyrrole, in complex prokaryotic signal transduction systems.

1. Introduction

Globins constitute an ancient superfamily of proteins, found in all domains of life and exhibiting enormous structural and functional diversity [1–3]. Because of their abundance, size, and ease of purification, globin proteins have been a prime model family for the study of protein structure, function, and evolution over the last century. Researcher Linus Pauling, in collaboration with others, such as Charles Coryell, was among the first whose research was centered on hemoglobin, the first ever identified globin protein, which utilizes a heme group to transport oxygen in the red blood cells of mammals [4,5]. Pauling and Coryell identified the structural change undergone by the hemoglobin subunits that is associated with the binding or loss of oxygen [6,7]. Later, Pauling, with his two co-workers—Robert Corey and Herman Branson, formulated a model of the hemoglobin structure in which amino acids are arranged so that they fold in a helical pattern [8]. This idea was further extended to explain the secondary structural elements of all proteins as being α-helices and β-sheets [9]. Not only did Pauling establish the fundamental chemistry behind protein structure, but his research also revealed that sickle-cell anemia is caused by a single amino acid mutation in hemoglobin [10]. This was the first ever proof of a disease as understood at a molecular level. Inspired by Pauling's groundbreaking research, Kendrew, Perutz, and their colleagues, ten years later, resolved the structure of myoglobin, which was the first protein structure resolved by means of X-ray crystallography [11]. Myoglobin is another heme-binding globin, well-known for its ability to bind and store oxygen in mammalian muscle cells [5]. Since then, studies of the globin protein have been crucial to the establishment of the homology concept and to the

[*] Corresponding author at: Department of Biology, College of Arts & Sciences, Saint Louis University, Saint Louis, MO 63105, United States.

E-mail address: dapeng.zhang@slu.edu (D. Zhang).

https://doi.org/10.1016/j.csbj.2021.12.022
2001-0370/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

early development of several bioinformatics and phylogenetic techniques [12–14].

In addition to serving as a prime model protein for biological scientists, the globin superfamily presents its own complex story of evolution and function. The traditional function of the globin is typified by mammalian hemoglobin and myoglobin, which adopt a characteristic structure composed of six core α-helices in a 3-over-3 configuration (referred to as the globin fold) [11,15]. The 3-over-3 α-helical structure is described as sandwiching a heme group which contains a central ferrous atom, enabling the oxygen binding ability of hemoglobin and myoglobin that is crucial to mammalian physiology [16]. However, numerous other heme-binding globin families have been identified and are found to display distinct functions [2,5,17,18]. For example, several plant globins, including symbiotic leghemoglobin, present in nitrogen-fixing root nodules of leguminous plants [19], and non-symbiotic phytoglobins [20–22], have functions ranging from oxygen buffering [23] and NO scavenging [24] to facilitating oxygen supply to developing tissues [25,26]. Aerotactic globin-coupled sensors, such as HemAT on the other hand, represent bacterial and archaeal globins that sense the level of diatomic oxygen in the cellular environment [27]. They mediate an aerophilic or aerophobic response by transmission of a signal through a linked signaling domain which interacts with the chemotaxis system [16,28,29]. Structural diversity within the globin superfamily is well represented by truncated globins (TrHbs), present in bacteria, unicellular eukaryotes, and higher plants [30,31]. TrHbs adopt a shortened α-helical structure with a 2-over-2 configuration, yet still maintain a recognizable globin fold [32]. Further, several new, distinct families have been discovered in animals, including neuroglobin [33], cytoglobin [34–36], globin X [37] , globin Y [38], globin E [39], and androglobin (Adgb) [40]. Except for Adgb, which is a chimeric permutated protein whose globin domain was split into two parts, the globin fold and its characteristic features are generally conserved among animal globins [40].

The globin repertoire further expanded with the identification of two non-heme binding globin families, including those identified in phycobiliproteins [41,42] and those in the RsbR family [43,44]. Both families adapt a globin fold, but they do not preserve the traditional distal and proximal histidine residues, which are responsible for holding the heme and gas molecules. Instead, they utilize other residues to configure an equivalent pocket for ligand binding. The phycocyanin family, used here to represent the globins found in different phycobiliproteins, such as phycoerythrin, phycocyanin, and allophycocyanin, is known to bind a linear tetrapyrrole chromophore, and functions as a light-harvesting molecule in cyanobacteria and red algae [41]. The RsbR globin family, on the other hand, potentially binds to a fatty acid [44] and functions as a sensor in a protein complex called a stressosome to receive environmental stress signals in *B. subtilis* [43] and in the process of sporulation inhibition in *B. anthracis* [44]. Although a clear evolutionary history remains elusive, the identification of non-heme binding families of globins has greatly expanded our understanding of the globin superfamily and has shed light on the unprecedented functional diversity within.

To gain insight into globin function and evolutionary history, we leveraged the most up-to-date genomic resources. Our exploration led us to identify a previously unrecognized globin family which we refer to as photoglobin. Through profile-based comparisons, clustering analysis, and structure prediction using AlphaFold2 [45], we show that photoglobin is related to phycocyanin and characterized by a comparable ligand-binding pocket, indicating that it might also bind a linear tetrapyrrole chromophore. Based on the contextual information gleaned from both domain architecture and gene-neighborhood analyses, we observe a strong association between the photoglobin, the B12-binding light sensor, and many other typical domains in prokaryotic signal transduction systems. Thus, we propose that the novel photoglobin family defines a new sensory mechanism in prokaryotic signal transduction, which, by means of a linear tetrapyrrole, transduces light signals to regulate photoglobin-associated enzymatic domains and transcription factors.

2. Materials and methods

2.1. Protein Domain-Centric computational concept

Our study features a protein domain-centric concept [46–48] in order to enhance the performance of the computational analysis. Proteins are typically comprised of one or multiple domains, which are the basic structural and functional modules of proteins and also evolutionary elements as their diversification and recombination delineate the diversity of the proteins [46]. Therefore, in order to avoid the unpredictable influences of multiple domains, we only applied discrete regions of protein domains, not full-length proteins, as the basic components in all computational analyses conducted in this study, such as sequence searches, comparisons, predictions and modeling. Further, by dissecting proteins into many conserved modules, we were able to accurately infer the functions of domains and collectively synthesize the functionality of proteins based on their domain architecture.

2.2. Homologous sequence searches and remote relationship detection

We utilized two computational strategies for homologous sequence searches and remote relationship detection. The first was an iterated profile-based method which includes the utilization of PSI-BLAST [49] and JACKHMMER [50] programs. PSI-BLAST is based on position-specific score matrix (PSSM) in which a new PSSM is generated and used as both a score system and a source of queries to search sequence databases. JACKHMMER is an equivalent program which generates a Hidden Markov Model (HMM) to enable sensitive sequence searches. With these two programs, we comprehensively collected homologs from the NCBI non-redundant (NR) sequence database for several known representative globin domain families, including heme-binding versions such as hemoglobin/cytoglobin [35,51], neuroglobin [33], androglobin [40], bacterial flavohemoglobin [52,53], protoglobin [54], and truncated globins [31,32,55], and the non-heme-binding versions, RsbR [43,44] and phycocyanin [41]. For most searches, a cut-off e-value of 0.01 was used to assess significance. In each iteration, the newly detected sequences that had e-values lower than the above cutoff were examined for being false positives and the search was continued with the same e-value threshold only if the profile was uncorrupted.

The second strategy was the profile-profile comparison method which is implemented in the HHpred program [56]. HHpred detects remote relationships between domains by comparing a profile HMM constructed from a PSI-BLAST search and a pre-computed library of profile HMMs compiled from domains recovered in the PDB database [56]. The significance was evaluated by both *E*-values (cut-off: 0.01) and probabilities (>95%). We also evaluated the significance by comparing the conserved motifs and the core of the fold to prevent false positive detections.

2.3. Sequence similarity-based clustering

For the sequences collected from the previous series of searches, similarity based clustering was performed to eliminate highly similar sequences by using CD-HIT, based on various sequence identity thresholds (0.4–0.7), and the BLASTCLUST program which adjusts the length of aligned regions and bit-score density threshold empirically (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The revised collection of the representative sequences from all major globin families was subjected to all-against-all BLASTP comparisons, and clustered based on the sequence similarity using the force-directed graph drawing (Fruchterman and Reingold) algorithm [57] implemented in the CLANS program [58].

2.4. Multiple sequence alignment and conservation analysis

Multiple sequence alignments (MSAs) were built for photoglobin and for other known globin families using PROMALS3D [59] and KALIGN [60]. For each globin family's alignment, careful manual adjustments were also performed to avoid introducing gaps into the sequences where consensus secondary structures were occupied and to maintain preservation of the typical globin structural arrangement and conservation of the proximal and distal residues involved in ligand binding. We further aligned representative globin domains with each other, to create a 'super alignment', based on structural superimpositions, profile-profile comparisons performed by HHpred [56], and secondary structure information predicted by the Jnet (Joint Network) program [61].

The conservation pattern of the MSA was computed by generating consensus residues based on different categories of residue properties [62]. The alignment was colored using an in-house alignment visualization program written in Perl and further modified using Adobe Illustrator.

2.5. Protein structure prediction, comparison and analysis

The protein structure prediction was conducted by using the recently developed deep learning system, AlphaFold2 [45]. According to the benchmark, AlphaFold2 generates a structure with atomic accuracy even where no similar structure is known [45]. Structure comparisons were conducted by using the DALI program, which generates an optimal pairwise structural alignment based on the similarity of local patterns extracted from contact maps [63]. Other structural analysis operations were conducted using the molecular visualization program PyMOL [64].

2.6. Guilt-by-association analysis of genomic contextual information

We utilized two computational guilt-by-association analysis strategies to infer the functional association of proteins and protein domains. The first one was a protein domain architecture analysis to identify the conserved domain fusion. Such domain fusion is typically a result of a gene fusion event during which interacting proteins or domains fuse together, thus facilitating effective functionality. Therefore, domain fusion is a strong indicator of a functional association (either direct or indirect interaction) between domains [46]. We detected protein domains using the HMMER package which searches against the Pfam database and a local collection of profile HMMs. For a given protein sequence or sequence region which did not match any existing domains, we conducted a series of sequence analysis steps including homologous sequence search, multiple sequence alignment, and profile-profile search to explore any potential new domains. This allowed for an initial functional annotation of proteins of interests based on domain architectures. The conserved domain fusion was then determined by frequency of the architecture and preservation in multiple bacterial lineages.

The second guilt-by-association strategy is the gene neighborhood analysis. It is based on the rationale that functionally linked genes in bacterial and archaeal genomes are often organized into operons or stay in proximity [47,65,66]. Such organization facilitates transcriptional regulation in prokaryotes since the genes in the same operon can be regulated by a single promotor. Thus,

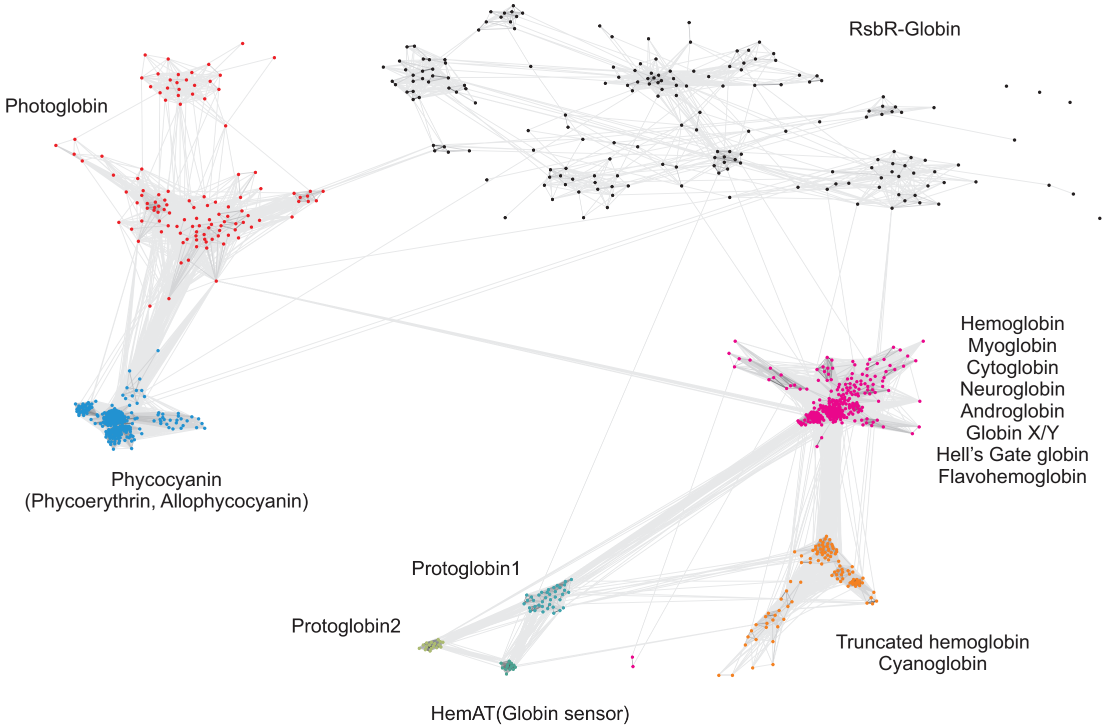
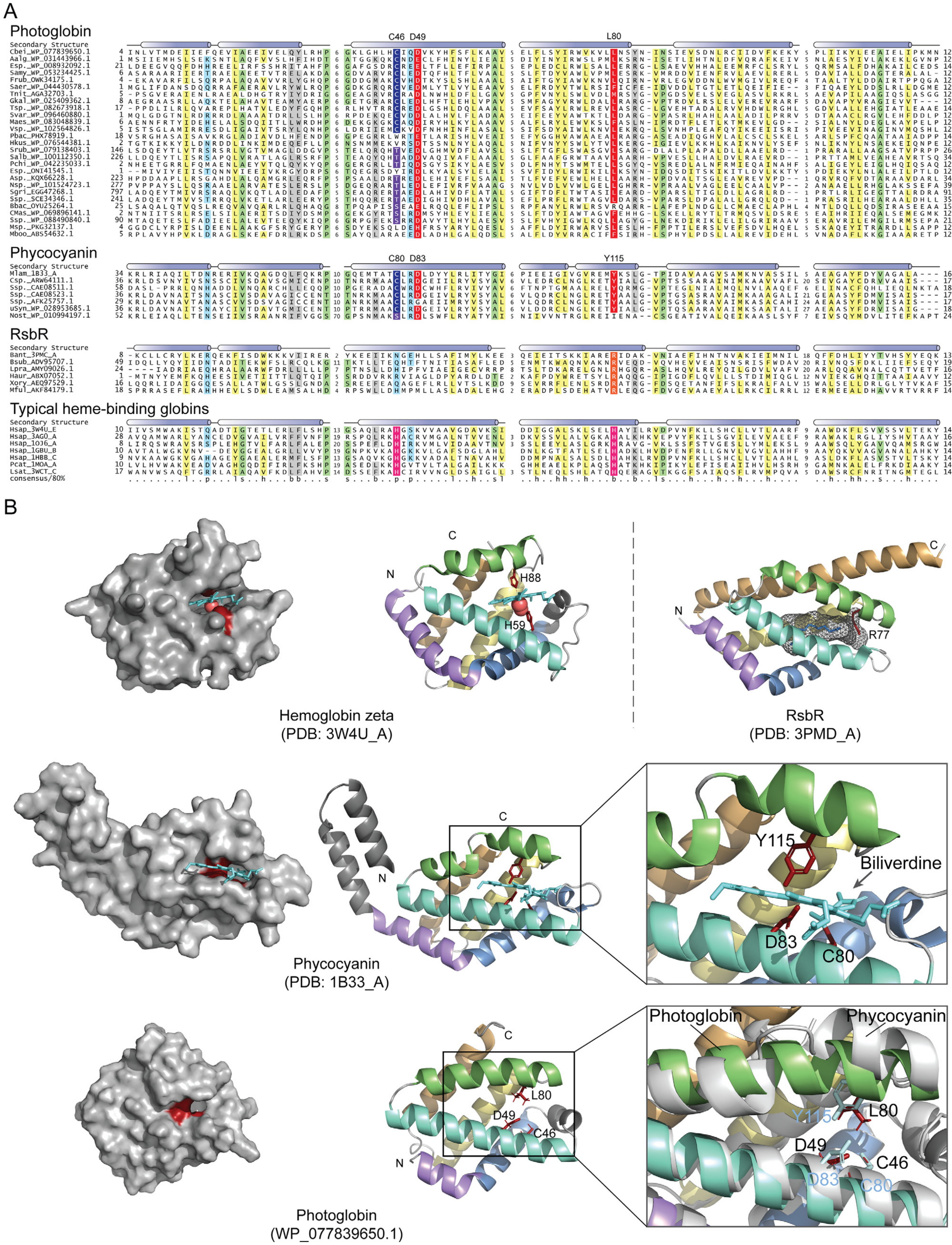


Fig. 1. Identification of a novel globin family by a clustering analysis of the globin sequence space. Sequence similarity between the globin domain sequences was detected by the BLAST program. Each node corresponds to a globin domain sequence. Straight lines indicate significant high-scoring segment pairs (HSPs) detected by all-against-all BLASTP searches with scoring matrix BLOSUM62 and an e-value cutoff of 0.005. The graph was generated by CLANS, which uses the Fruchterman and Reingold graph drawing algorithm.


A
Photoglobin
Secondary Structure
C46 D49
L80
Cbei_WP_077839650.1
Aalg_WP_031443966.1
Esp._WP_008932092.1
Samy_WP_053234425.1
Frub_OWK34175.1
Saer_WP_044430578.1
Tnit_AGA32703.1
Gkal_WP_025409362.1
Tsp._WP_082673918.1
Svar_WP_096460880.1
Maes_WP_083048839.1
Vsp._WP_102564826.1
Pbac_PHX78919.1
Hkus_WP_076544381.1
Srub_WP_079138403.1
Salb_WP_100112350.1
Pchi_WP_042235033.1
Esp._ONI41545.1
Asp._KQX66228.1
Nsp._WP_101524723.1
Sgri_EGG47268.1
Ssp._SCE34346.1
Bbac_OYU25264.1
Ssp._AFK25757.1
CMas_WP_069896141.1
Ssp._WP_088490840.1
Msp._PKG32137.1
Mboo_ABS54632.1
Phycocyanin
Secondary Structure
C80 D83
Y115
Mlam_1B33_A
Csp._ARW64111.1
Ssp._CAE08511.1
Ssp._CAE08523.1
Ssp._AFK25757.1
uSyn_WP_028953685.1
Nost_WP_010994197.1
RsbR
Secondary Structure
Bant_3PMC_A
Bsub_ADV95707.1
Lpra_AMY09026.1
Haur_ABX07052.1
Xory_AEQ97529.1
Mful_AKF84179.1
Typical heme-binding globins
Secondary Structure
Hsap_3W4U_E
Hsap_3AG0_A
Hsap_1O36_A
Hsap_1GBU_B
Hsap_1HBB_C
Pcat_1MOA_A
Lsat_3WCT_C
consensus/80%
B
C
N
H88
H59
Hemoglobin zeta
(PDB: 3W4U_A)
C
N
R77
RsbR
(PDB: 3PMD_A)
C
N
Phycocyanin
(PDB: 1B33_A)
Y115
Biliverdine
D83
C80
C
L80
D49
C46
N
Photoglobin
(WP_077839650.1)
Photoglobin
Phycocyanin
Y115
L80
D49
C46
D83
C80

membership in the same operon can be utilized as evidence for functional association. We have developed a dedicated pipeline to systematically identify such gene neighborhoods. Specifically, a custom Perl script is used to extract gene neighbors (usually 7–10 genes) on either side of the query (all photoglobin gene homologs) from the PTT file (downloadable from the NCBI ftp site) or the Genbank file in the case of whole genome sequences available. The protein sequences of all gene neighbors were then clustered using the BLASTCLUST program. Each cluster of homologous proteins was annotated based on the domain architecture analysis using the HMMER package. We determined the conserved co-occurring genes based on 1) the frequency of common domains identified in the neighborhood, 2) the presence of such co-occurring genes in at least two phylogenetically distinct lineages ("phylum" in NCBI Taxonomy database), and 3) complete conservation in a particular lineage ("phylum").

3. Results and discussion

3.1. Sensitive sequence searches and clustering analysis uncover a novel family of globins

We first recovered the photoglobin domain family while collecting homologous sequences of the phycocyanin family using PSI-BLAST searches against the NCBI-non-redundant (NR) sequence database. In this search, we used a phycocyanin beta subunit from the cyanobacteria *Fischerella* as the query (accession number: WP_026736665.1); in the 4th iteration, we retrieved about 24 thousand sequences which contained the phycocyanin domains. The majority of the sequences (>97%) were either from cyanobacteria or red algae, which are known to use phycocyanin-based protein complexes, phycobilisomes, for light harvesting [67,68]. Interestingly, we noticed some sequences named 'cobalamin B12-binding domain-containing proteins' which started to appear as significant hits (e-values around 0.003), but with low sequence identify (below 20%) to the query phycocyanin. According to the Pfam domain annotations, the corresponding regions of most of these proteins were unclassified. To understand the identity of the potential protein module found in these B12-binding domain proteins, we performed a PSI-BLAST search with successive iterations using one of the unclassified sequences from *Streptomyces* (accession number: WP_081221060.1, residues 224–350) as the query. The searches identified many homologues from a broad range of species of both bacteria and archaea, a phyletic distribution distinct from the one observed in the case of phycocyanin. In addition to many sequences annotated as cobalamin B12-binding domain-containing proteins (e.g., WP_081221060.1), other sequences containing this potential module include MerR family transcriptional regulators (e.g., RPI92725.1), methanogenic corrinoid proteins MtbC1 (e.g., RDI21513.1), biliverdin-producing heme oxygenase (e.g., WP_045221611.1), twin-arginine translocation pathway signal proteins (e.g., ORV98546.1) and many hypothetical proteins. None of the sequences have been previously reported to contain any known globin domain. Also, HMMER scans against Pfam domain profiles failed to detect any known domains for the majority of the sequences, indicating they might represent a new globin family.

We next initiated several subsequent PSI-BLAST searches with distinct bacterial and archaeal sequences as the respective query sequences to achieve maximal coverage in terms of detection of this potential globin domain. As a result, we identified about 5975 homologues from the NCBI-NR database released on June 7, 2021. We also comprehensively collected homologues from the NCBI-NR database for several known globin domain families, including heme-binding versions, hemoglobin/cytoglobin [35,51], neuroglobin [33], androglobin [40], bacterial flavohemoglobin [52,53], protoglobin [54], truncated globins [31,32,55] , and non-heme-binding versions, RsbR [43,44] and phycocyanin [41]. For each globin family, we removed the highly similar sequences by CD-HIT [69].

Next, all representatives of known globin domain families were combined with the above-collected novel globin domain sequences, subjected to all-against-all BLASTP comparisons, and clustered based on the sequence similarity using the force-directed graph drawing (Fruchterman and Reingold) algorithm [57] implemented in the CLANS program [58]. As shown in Fig. 1, the sequences, presented as nodes in this two-dimensional graph, are linked and clustered by edges, denoting the detected sequence similarity when they are from the same or closely related globin domain families. For example, sequences of several vertebrate globin families such as hemoglobin/cytoglobin, neuroglobin, androglobin, and many bacterial sequences of flavohemoglobin and Hell's Gate globin [70] appear to be clustered as one group. Sequences of the protoglobin and globin-coupled sensor/HemAT family are clustered into three related groups, whereas sequences belonging to the truncated globins, including cyanoglobin [55], are clustered together. Furthermore, the phycocyanin sequences are clustered into a dense group, while the RsbR-related sequences [43] and the fatty-acid binding globin sensor [44] compose a large dispersed cluster. By contrast, the novel globin sequences (labeled as photoglobin) present themselves as a large, self-connected cluster in this globin sequence space, exclusive of all other known globins, and only display limited linkages to some sequences of other families, such as phycocyanin. Thus, these results suggest that the collected sequences might represent a novel globin domain family that has not been recognized before. Although our decision to refer to the unclassified globin domain as the photoglobin came after additional, extensive sequence, structural and genomic analyses, to maintain clarity throughout this paper, we will refer to this glo-

Fig. 2. (A) Multiple sequence alignment (MSA) between photoglobin, phycocyanin, RsbR-globin, and other typical heme-binding globin families. The secondary structure of each globin family is shown above the MSA and the consensus information is shown below the MSA, where h stands for hydrophobic residues, l for aliphatic residues, s for small residues, b for big residues and p for polar residues. The conserved residues involved in ligand-binding are highlighted in white font with a different background color in each family, which are shown in dark blue/purple and red in photoglobin and phycocyanin, orange in RsbR-globin and hot pink in other typical heme-binding globins. (B) Structural comparison of a typical heme-binding globin (hemoglobin), RsbR-globin, phycocyanin, and the AlphaFold2 predicted model of photoglobin. Both surface view and cartoon view are presented. The six-core α-helices of these globin structures are colored differently: first α-helix in purple, second in blue, third in cyan, forth in green, fifth in yellow and sixth in orange. The conserved residues involved in binding specific ligands are colored in red. The bottom-right figure shows a structural alignment between phycocyanin and photoglobin domains. Species abbreviations used in the figures are: Aant: Amycolatopsis antarctica; Achi: Actinomycetospora chiangmaiensis; Aspr: Angustibacter sp.; Athe: Anaerolinea thermophila; Bant: Bacillus anthracis; Bbac: Burkholderiales bacterium; Bspp: Bacillus sp.; Bsub: Bacillus subtilis; Chon: Chondria sp. (in: Rhodophyta); Espp: Ectothiorhodospira sp.; Esps: Epulopiscium sp.; Haur: Herpetosiphon aurantiacus; Hkus: Halanaerobium kushneri; Hsap: Homo sapiens; Lpra: Luteitalea pratensis; Lsat: Lamellibrachia satsuma; Mass: Candidatus Marispirochaeta associata; Mbol: Melittangium boletus; Mboo: Methanoregula boonei; Meth: Methanoregula sp.; Mful: Myxococcus fulvus; Mlam: Mastigocladus laminosus; Nost: Nostocaceae; Nsp7: Nocardioides sp.; Pbac: Planctomycetaceae bacterium; Pcat: Physeter catodon; Pchi: Paenibacillus chitinolyticus; Ppar: Pedosphaera parvula; Psph: Proteocatella sphenisci; Salb: Streptomyces alboflavus; Sgri: Streptomyces griseoaurantiacus; Srub: Streptomyces rubidus; Sspo: Saccharibacillus sp.; Ssps: Streptomyces sp.; Sspw: Synechococcus sp.; Svar: Sulfurifustis variabilis; Svir: Saccharomonospora viridis; Xory: Xanthomonas oryzae. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

bin domain and globin family as photoglobin from this point forward.

3.2. Sequence and structural analysis of the photoglobin family

During the PSI-BLAST searches of photoglobin homologs, we observed that several phycocyanin globin sequences were included as matched hits (either significant or insignificant) starting in the iterations from 3 to 8 (end of our search). Further, a sensitive HMM profile-profile comparison against PDB structures implemented in the HHpred server returned a significant match of the photoglobin domain to several phycocyanin globin structures (e.g. Phycocyanin Alpha Chain, PDB: 2VML_G, probability 99.84% of profile-profile match; Allophycocyanin beta-18 subunit apoprotein, PDB: 6HRK_C, probability 99.84% of profile-profile match). Additionally, our CLANS clustering analysis showed a good number of linkages between the photoglobin and phycocyanin domain sequences. Thus, these sequence/profile comparisons indicate that the photoglobin domain might be related to the phycocynanin globins.

To further understand this potential homologous relationship, we conducted sequence and structural analysis on this photoglobin domain with a comparison to several known globin families including the non-heme binding globins, phycocyanin and RsbR, and typical heme-binding globins from animals (Fig. 2). We found that, among these known diverse globin families, the core of the globin fold is composed of six α-helices in a 3-over-3 configuration (which we will denote as helices 1–6). This is true despite the fact that some modifications exist, as is the case for the hemoglobin with PDB structure 3W4U_A, shown in Fig. 2, in which two additional short alpha helices are present in the position of the loops between helices 2–3 and 5–6; modifications also exist in the case of truncated globins (PDB: 1DLY_A), in which helix 1 is absent and the fourth helix takes a loop-like form with a single-turn helix, causing the globin to adopt a 2-over-2 configuration [71]. Despite these modifications, the ligand-binding pocket located between the third and fourth helices appears to be conserved across the families. Through structure-guided sequence alignment, we found that, while the ligands (cofactors) are different between heme-binding and non-heme binding globin families, their ligand-binding residues are evolutionarily related. For example, the typical heme-binding globin families characteristically utilize two conserved residues (a distal histidine and a proximal histidine) to hold a heme group and/or a gas molecule, such as oxygen in place (Fig. 2A). Within the phycocyanin family, the ligand, a linear tetrapyrrole chromophore (bilin), is held through a covalent linkage via a cysteine (C80, which is structurally analogous to the distal histidine) and noncovalent linkages via an aspartic acid (D83) and a hydrophobic residue (usually tyrosine, Y115; structurally analogous to the proximal histidine) (Fig. 2B); in the case of the RsbR family, the ligand, a fatty acid, is bound by a conserved arginine (R), which is again structurally analogous to the proximal histidine (Fig. 2B). Thus, these family-specific residues seem to have defined the ligand-binding specificity of the globins.

When examining the sequence features of the photoglobin family (Fig. 2A), we found that photoglobin displays a secondary structural configuration like that of the traditional globin with six core α-helices, although its sequence identity to other globins was low. Two highly conserved positions, an aspartic acid (D49) on the third helix and a hydrophobic residue (typically Leucine L80) on the fourth helix, can be identified in the majority of photoglobins, which appear to be well aligned with the phycocyanin-specific residues D83 and Y115. Although the residue likely to be involved in distally binding the ligand is not conserved among photoglobin sequences, we did observe that some of the photogolobins preserve a Cys in the same position as the conserved ligand-binding C80 in phycocyanin (Fig. 2A). Importantly, many of the photoglobins that do not preserve a cysteine in the distal ligand-binding position have a serine or threonine residue instead. Given that both Ser and Thr have a hydroxyl group (–OH), which is chemically similar to the sulfhydryl group (-SH) of Cys, in that both are nucleophollic, it is conceivable that Ser and Thr might function similarly in providing a covalent linkage with the ligand.

To further support our sequence analysis, we utilized the newly developed AlphaFold2 [45] to predict the potential structure of a photoglobin representative (WP_077839650.1, residues 1–130). The structure of the photoglobin indeed adapts a typical globin fold with a potential ligand-binding pocket formed between α-helices 3 and 4. Further, all three potential ligand-binding residues (C46, D49, L80) are located on the inner surface of the pocket and exhibit a similar configuration as the ligand-binding residues in phycocyanin (Fig. 2B). Given the observed similarity between the ligand-binding pockets of photoglobin and phycocyanin, we propose that the new photoglobin domain might bind a ligand or a series of ligands, which are structurally similar to the linear tetrapyrroles bound by phycocyanin. We also predict that the photoglobin domains respectively utilize not only cysteine, but also serine or threonine residues located in the distal-binding position to covalently link their ligands, and that the residue variation in the distal-binding position might be attributed to photoglobin's ability to bind different types of linear tetrapyrroles.

3.3. Photoglobin is distinct from phycocyanin in both structure and phyletic distribution

Despite the many similarities between photoglobin and phycocyanin in both structure and ligand-binding configurations, there are still two major differences between them. First, the photoglobin domain is missing the two unique α-helices that the phycocyanin has at its N-terminus ahead of the first shared helix of the globin fold (Fig. 2B). This N-terminal α-helix pair is critical for interaction of phycocyanin monomers which will further aggregate to form ring-shaped trimers and hexamers [72], the basic unit of the pigment protein complex called the phycobilisome [67]. Therefore, photoglobin is unlikely to form a higher-order structure complex like phycocyanin does. The second major difference between photoglobin and phycocyanin is that phycocyanin is predominantly present in cyanobacteria and eukaryotic red algae (rhodophyta), which use phycocyanin-based phycobilisomes as accessory pigments to chlorophyll [67]. Photoglobin, on the other hand, has a broader distribution, as it is present in many different bacterial lineages, including actinobacteria, firmicutes, proteobacteria, and chloroflexi, as well as in archaea (Fig. 3), which are not known to utilize light for energy. These differences suggest that photoglobin might have a function that is unlike that of phycocyanin's function as a light-harvesting pigment.

3.4. Contextual information reveals that photoglobin is involved in prokaryotic signal transduction systems

To explore the potential function of photoglobin, we utilized a computational "guilt-by-association" strategy of mining contextual information, such as the domain architectures of the photoglobin proteins and their gene neighborhoods, on bacterial and archaeal genomes. The principle of this strategy is that protein domains that interact or are functionally linked usually display a conserved association in either domain architectures or gene neighborhoods across different species. By extracting such information about these associations (contextual information), we can infer the functional context of the photoglobin. The results of our analysis of the photoglobin domain architecture (Fig. 4) reveal that, in most cases, the photoglobin domain is found fused to the so-

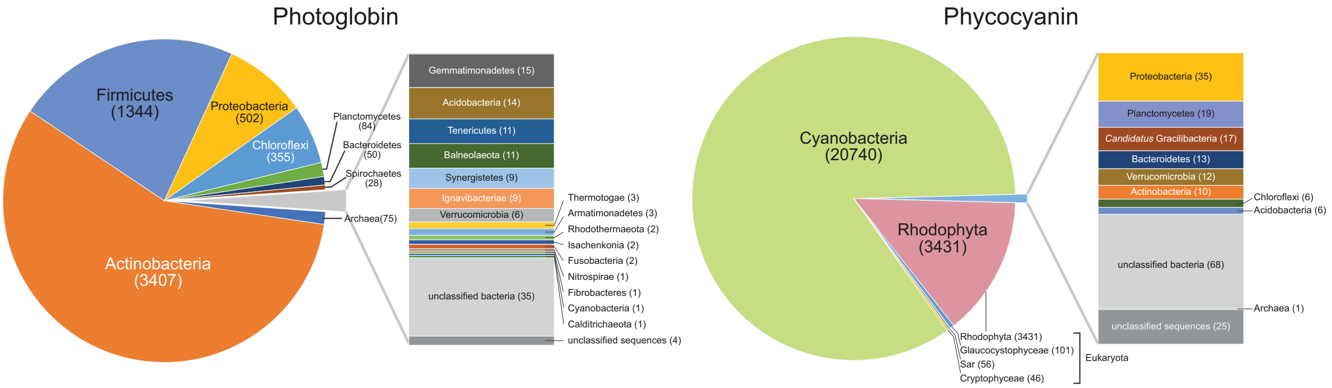


Fig. 3. Taxonomic Distribution of the photoglobin family with a comparison with the Phycocyanin family.

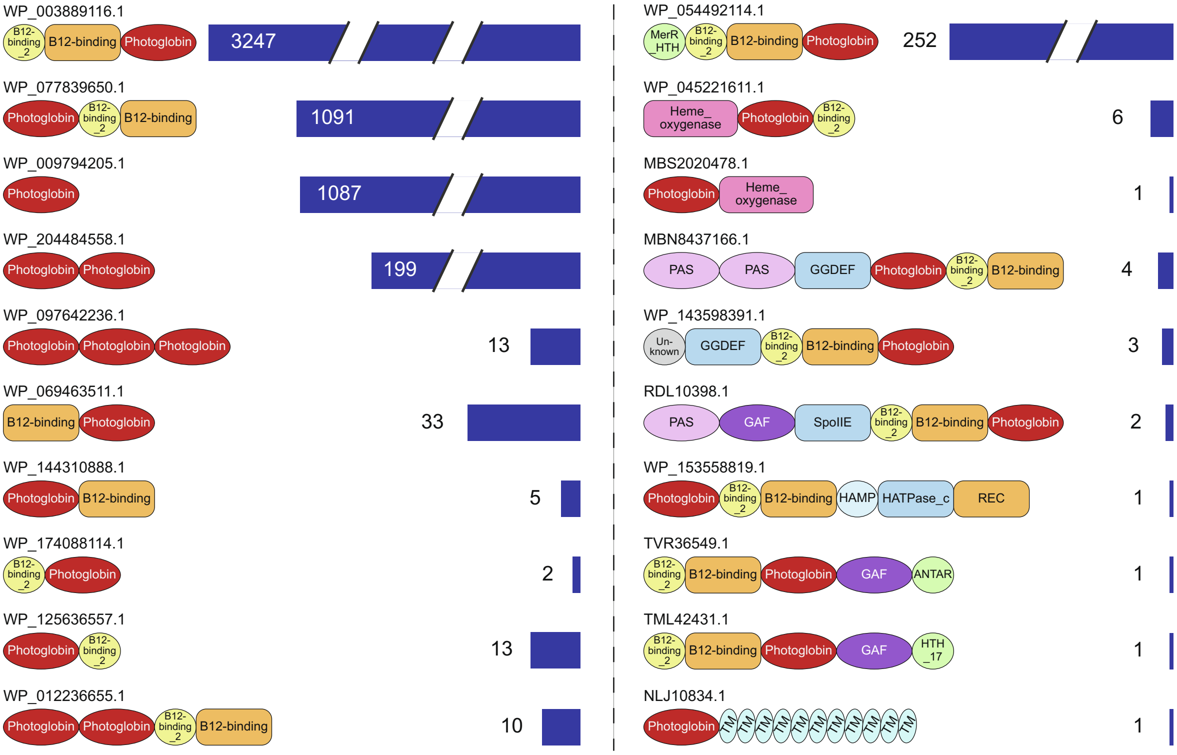

WP_003889116.1
B12-binding_2 B12-binding Photoglobin
3247
WP_077839650.1
Photoglobin B12-binding_2 B12-binding
1091
WP_009794205.1
Photoglobin
1087
WP_204484558.1
Photoglobin Photoglobin
199
WP_097642236.1
Photoglobin Photoglobin Photoglobin
13
WP_069463511.1
B12-binding Photoglobin
33
WP_144310888.1
Photoglobin B12-binding
5
WP_174088114.1
B12-binding_2 Photoglobin
2
WP_125636557.1
Photoglobin B12-binding_2
13
WP_012236655.1
Photoglobin Photoglobin B12-binding_2 B12-binding
10
WP_054492114.1
MerR_HTH B12-binding_2 B12-binding Photoglobin
252
WP_045221611.1
Heme_oxygenase Photoglobin B12-binding_2
6
MBS2020478.1
Photoglobin Heme_oxygenase
1
MBN8437166.1
PAS PAS GGDEF Photoglobin B12-binding_2 B12-binding
4
WP_143598391.1
Un-known GGDEF B12-binding_2 B12-binding Photoglobin
3
RDL10398.1
PAS GAF SpoIIE B12-binding_2 B12-binding Photoglobin
2
WP_153558819.1
Photoglobin B12-binding_2 B12-binding HAMP HATPase_c REC
1
TVR36549.1
B12-binding_2 B12-binding Photoglobin GAF ANTAR
1
TML42431.1
B12-binding_2 B12-binding Photoglobin GAF HTH_17
1
NLJ10834.1
Photoglobin TM TM TM TM TM TM TM TM TM TM
1

Fig. 4. Domain architectures of representative photoglobin domain-containing proteins. The proteins are labeled with their NCBI accession numbers and are not drawn to scale. The frequency of each domain architecture is shown to the right of the architecture cartoon with the length of the blue bar corresponding to the explicit frequency. The Pfam IDs of the included domains are: B12-binding, PF02310; B12-binding_2, PF02607; MerR-HTH, PF13411; Heme_oxygenase, PF01126; PAS, CL0183; GAF, CL0161 ; GGDEF, PF00990; SpoIIE, PF07228; HAMP, PF00672; HATPase_c, PF02518; REC, PF00072 (Response_reg); ANTAR, PF03861; HTH_17, PF12728. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

called B12-binding domain_2 (Pfam ID: PF02607) and B12-binding (Pfam ID: PF02310) domains, located consecutively either before or after the photoglobin domain. The B12-binding domain is known to utilize either the vitamin B_{12} derivative adenosylcobalamin ($AdoB_{12}$) to sense light and regulate transcription factor activation or to utilize methylcobalamin (MeB_{12}) to facilitate transfer of a methyl group in methionine synthase; the B12-binding_2 domain is a four-helical bundle which caps the upper face of $AdoB_{12}$ or MeB_{12} in their respective reactions [73,74]. Additionally, the photoglobin domain is frequently found as a stand-alone domain or multiply linked. In several other situations (Fig. 4), photoglobin is further coupled with, in addition to the B12-binding_2 and B12-binding domains, several domains known to be involved in prokaryotic signal transduction systems, including PAS, GAF, GGDEF, Stage II sporulation protein E (SpoIIE), histidine kinase, and transcription factors such as MerR-HTH, ANTAR, and HTH_17 (Fig. 4). This suggests that the photoglobin domain might be functionally linked to signaling systems.

To gain a more complete understanding of the potential functional relevance of photoglobin, we conducted an operonic analysis on the 5975 photoglobin genes identified from our PSI-BLAST searches. Fig. 5 shows the representative gene neighborhood associations of the photoglobin-containing loci in both bacteria and archaea. None of the genomic loci have been experimentally studied. Even though their gene composition is highly divergent, we were able to identify two striking themes. First, standalone photoglobin domains are almost always located adjacent to the B12-binding_2 and B12-binding domains on their respective operons. Second, in almost all instances, the photoglobin genes were further associated with genes of prokaryotic signal transduction systems,

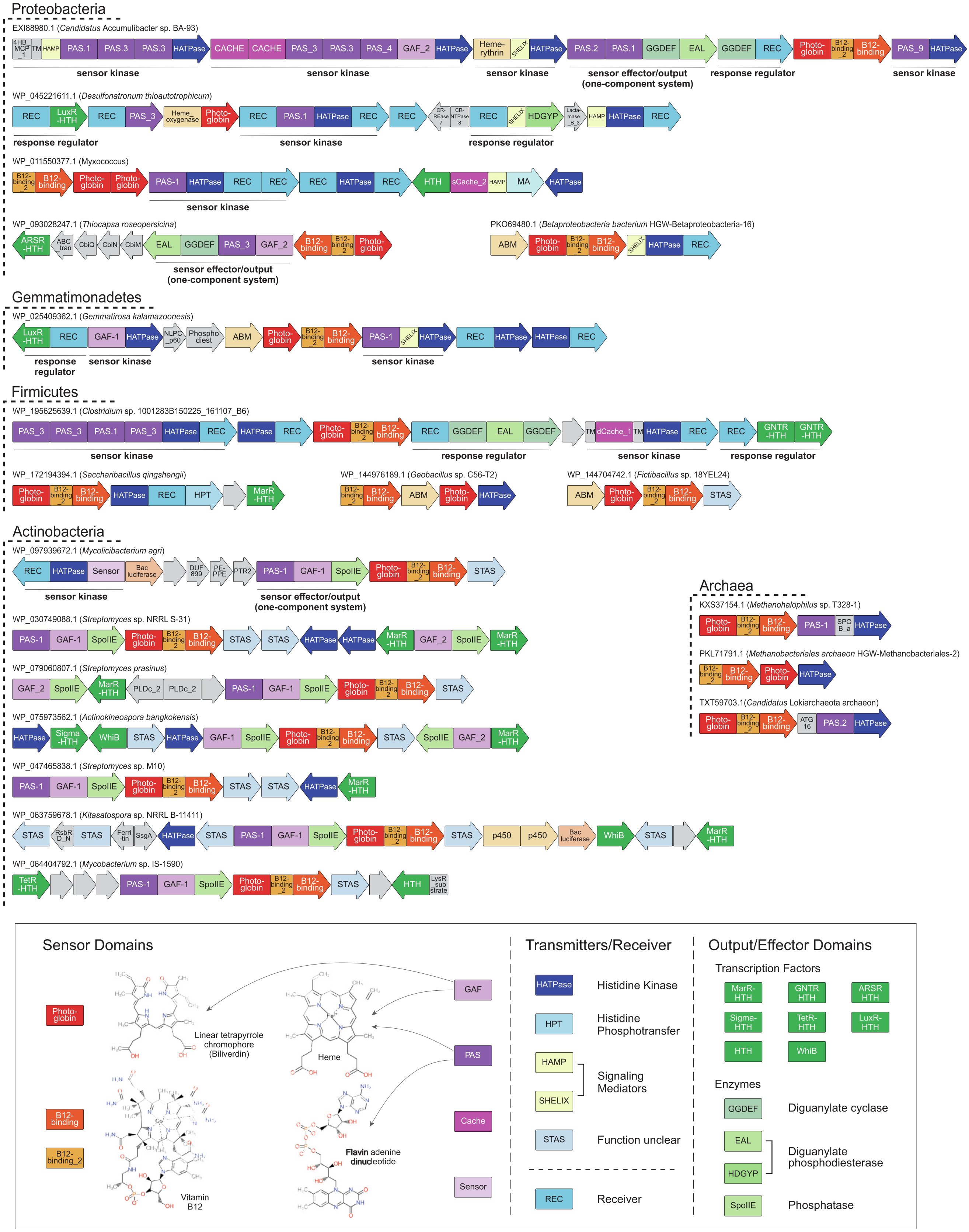


Fig. 5. Representative operonic structures of the photoglobin-containing loci. Each gene is presented as a block arrow, in which the major domains of the encoded protein are separately shown as rectangular segments. The direction of the arrow shows the direction of the transcription. The loci are labelled by the photoglobin-containing protein accession numbers and species names in bracket. The included domains can be classified into three major categories based on their function in the signal transduction systems: sensor domains, transmitters/receiver, and output/effector domains (transcription factors and enzymes). The three major groups of protein components are labeled on selected operons as sensor kinases, response regulators, and sensor effectors (details in main text). The chemical structures of the typical ligands of several sensors are also shown.

mostly typified by two-component and one-component systems. Two-component and one-component systems represent the primary modes of signal transduction in bacteria and archaea to detect and respond to environmental stresses, including pH, light, gas molecules, and ion concentrations [75,76]. They typically utilize many distinct domains of three functional categories — sensors, transmitters/receiver, and output/effector domains — to construct a variety of sensor kinases and response regulators (in two-component systems) or sensor effectors (in one-component systems) [76]. In a prototypical model, a sensor domain, often periplasmic, converts an environmental signal to activate a histidine kinase, via autophosphorylation of a conserved histidine [77]. The activated histidine kinase acts as a transmitter, transferring the phosphoryl group to a conserved aspartate found on the receiver domain of a response regulator, whose output domain is typically a transcriptional regulator [78]; this in turn modulates the transcription factor's DNA binding activity, thereby regulating gene expression [75].

Here, from the photoglobin-containing operons, we have recovered and annotated many typical domains of these systems, which fall into the three main functional categories— sensors, transmitters/receiver, and output/effector domains. First, this includes distinct versions of several sensor domains, namely PAS, GAF, and Cache, which are the predominant sensor modules in prokaryotic signal transduction systems [76,79–81]. Second, in addition to the primary histidine kinase (labelled HATPase), several other transmitters were recovered, including the histidine phosphotransfer domain (Hpt) [82], HAMP [83,84], SHELIX [85] and the STAS domain [86,87], which displays a ubiquitous presence in actinobacteria. The receiver domain (REC) is found fused to both transmitter and output domains throughout the photoglobin operons. Third, several types of output/effector domains were recovered, including both transcription factors and enzymatic domains. The transcription factors include a diverse group of DNA-binding transcription factors adapting the helix-turn-helix (HTH) fold [88], along with the four-helix bundle transcription factor WhiB, present in actinobacteria [89]. Enzymatic output/effector domains include the GGDEF diguanylate cyclase and EAL/HD-GYP diguanylate phosphodiesterases, which catalyze the formation and breaking down of cyclic nucleotide second messenger cGMP, respectively [90–92]. Additionally, the enzymatic output serine phosphatase domain, SpoIIE, displays a marked presence in actinobacteria. Though not commonly studied in the context of one- or two-component signaling, SpoIIE is known to de-phosphorylate the sulfate transporter and anti-sigma factor antagonist (STAS) domain containing protein, SpoIIAA, in the process of regulation of a sporulation-specific transcription factor, σ^F [93–95]. In addition to these three main groups, we also identified several oxygenases, including bacterial luciferase, ABM monooxygenase, hemeoxygenase and p450 which might be responsible for sensing and degrading respective chemicals such as flavin, antibiotics, and heme [96–98].

Based the domain annotations on the photoglobin-containing operons, we were able to predict the potential role of the protein components and their organizations/interactions in the respective systems (Fig. 5). First, we recovered many distinct versions of sensor kinases, which are featured by containing one or multiple sensor domains along with the histidine kinase domain. Second, we recovered many response regulators, which contain the receiver domain and one or multiple output/effector domains. Third, we recovered many sensor effectors (of one-component systems) which are proteins in which the sensor and effector/output domain are directly fused. Importantly, we found that many operons contain multiple sets of sensor kinases and response regulators, such as those containing EXI88980.1 in proteobacteria and WP_195625639.1 in firmicutes (Fig. 5), indicating that they might define novel signal transduction systems which involve multikinase signaling cascades integrating diverse stimuli [99]. Additionally, there are several atypical configurations. For instance, in the majority of operons in actinobacteria, the histidine kinase proteins and transcription factors occur as standalone proteins, while the PAS and GAF sensor domains are directly fused with the SpoIIE phosphatase domain. How the components of those systems are coordinating is unknown.

What is the functional role of the photoglobin proteins in these prokaryotic signal transduction systems? The photolgobin and B12-binding domains are almost always fused or coupled together at the level of protein organization (Fig. 4) or genomic association (Fig. 5), indicating that they function cooperatively. In some instances, they are found to be additionally fused to different effector/output domains, as shown by the domain architectures of WP_054492114 and TVR36549.1 (Fig. 4). This organization mimics that of other sensor effectors in one-component systems, suggesting that photoglobin and the B12-binding domains might function as sensors. Further, when analyzing the known sensor domains, we realized that a chemical foundation of these domains is the use of various tetrapyrroles or nucleotide derivatives to receive a light signal, chelate gas molecules, or to sense redox potential. For example, the PAS domain, the most ubiquitous sensor domain, has been shown to bind various ligands such as tetrapyrrole heme, flavin mononucleotide (FMN), flavin adenine dinucleotide (FAD), di-/tricarboxylate, and 4-hydroxycinnamic acid [100–102]. The GAF domain, another prevalent sensor, is known to sense light by utilizing heme [103] and several linear tetrapyrrole chromophores [104], a similar set of ligands that phycocyanin-globin binds to. Therefore, both the novel photoglobin and the B12-binding domains are functionally aligned with these sensors, as 1) photoglobin is predicted to bind a linear tetrapyrrole and 2) the B12-binding domain, although it is not commonly implicated in either one-component or two-component signaling, has been shown to rely on the tetrapyrrole adenosyl-B_{12} ($AdoB_{12}$) to mediate a light-induced regulation of transcription factors [74]. Thus, given the observed strong association between the photoglobin and many signal transduction domains, coupled with our earlier prediction of binding of a linear tetrapyrrole by photoglobin, we propose that photoglobin might act as a novel light sensor, which utilizes a linear tetrapyrrole (bilin) to convert a light signal to regulate associated enzymes and transcription factors in prokaryotic signal transduction systems.

3.5. Structural prediction reveals that photoglobin and B12-binding domains constitute a coupled photosensing system

As observed in the photoglobin-containing proteins and operons, prokaryotic signal transduction systems are typically characterized by a great diversity of organization in both protein domain architecture and genomic loci. In order to understand the potential interactions between these domains in these signaling systems, we next sought to apply the newly developed AlphaFold2 system [45] to predict the structures of several representative photoglobin-containing proteins (Fig. 6). The structures predicted by AlphaFold2 confirmed our annotations on the domain architectures of these proteins. Interestingly, even though many proteins have different domain composition and organization, in which signaling domains are combined and arranged differently in sequence, we found that the structural arrangement between three domains— photoglobin, B12-binding, and B12-binding_2—appears to be stable (Fig. 6). This indicates that the mode of interaction between the photoglobin and B12-binding domains is conserved during evolution. Such conservation might also indicate that these two light sensor domains might function together as a regulatory bundle. To be noted, B12-binding domain, via the $AdoB_{12}$ ligand,

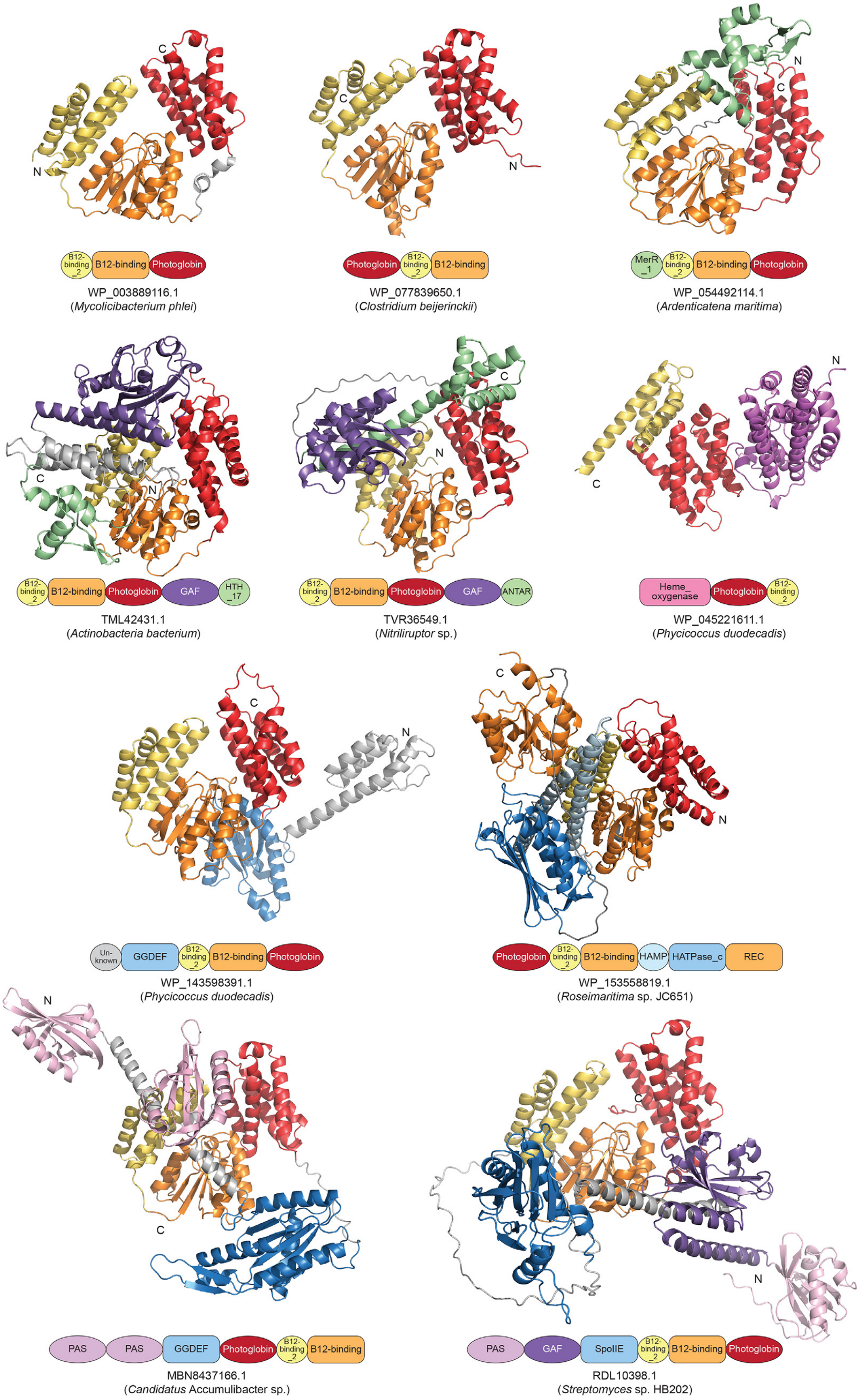


Fig. 6. AlphaFold2 predicted structures of representative photoglobin-containing proteins. Their domain architectures are shown below the structures. The color themes of the structures correspond to the that of their respective domain architectures.

can sense light over a spectral range from near-UV to visible light of wavelengths < 530 nm (corresponding to blue and green) with peak absorbance at 365 nm [74]. Photoglobin, on the other hand, is predicted to sense longer wavelengths (corresponding to orange and red) given its similarity to the phycobiliproteins (phycocyanin, phycoerythrin, allophycocyanin) which are covalently bound to linear tetrapyrrole chromophores including phycoerythrobilin (PEB), phycocyanobilin (CYC), and phycourobilin (PUB) [105,106]. Therefore, it is possible that the coupling of the photoglobin and B12-binding domains might facilitate a switch-like regulation of downstream signaling components in response to different wavelengths of visible light. This is likely the first case where we see two sensor domains coupled together in the prokaryotic signal transduction systems.

In addition to the conserved interactions between photoglobin and B12-binding domains, structural models also reveal several other contacts between the photoglobin/B12-binding domain hub and other signaling components, such as transcription factor domains (as seen in WP_054492114.1 and TVR36549.1, Fig. 6). There are several instances in which the histidine kinase domain or enzymatic effectors are distant to the photoglobin/B12-binding domain hub (as seen in WP_153558819.1 and MBN8437166.1 in Fig. 6), which could indicate that the sensor, upon activation by light, might allosterically modulate the downstream effectors or transcription factors, effectively switching the signal transduction pathway on and off.

4. Final remarks

In summary, by using a multi-faceted analysis of sequence, structure, and genomic organization, we have identified and made functional inferences for a new family of globins which we have named photoglobin for its potential role in light sensing. We show that photoglobin adopts a typical 3-over-3 globin fold and preserves a ligand-binding pocket like that of phycocyanin. Via the highly conserved distal cysteine, aspartate and proximal leucine residues, photoglobin likely binds a linear tetrapyrrole thought to enable light sensing ability. By extracting contextual information from protein domain architectures and gene neighborhoods, we found that photoglobin is almost always fused to or located adjacent to the B12-binding domains, which, together, are strongly associated with many typical domains of two-component signal transduction systems. By examining a series of structures of photoglobin-containing proteins modeled by AlphaFold2, we propose that photoglobin may act in conjunction with the B12-binding domains as a light-sensing regulatory bundle in prokaryotic signal transduction systems. Together, photoglobin and the B12-binding domains may regulate operonically-associated enzymes or transcription factors in response to different light conditions. We expect that our discovery of this novel globin family in complex signaling systems will encourage other experimental biologists to systematically study the functions of these proteins in microbial signal transduction, hopefully leading to applications of this novel type of photo-switch in future biotechnology.

CRediT authorship contribution statement

Theresa Schneider: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Yongjun Tan:** Methodology, Software, Formal analysis, Writing – review & editing, Visualization. **Huan Li:** Methodology, Software, Writing – review & editing. **Jonathan S. Fisher:** Conceptualization, Writing – review & editing. **Dapeng Zhang:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Saint Louis University Start-up fund.

References

[1] Vinogradov SN, Moens L. Diversity of globin function: enzymatic, transport, storage, and sensing. J Biol Chem 2008;283(14):8773–7.
[2] Hardison RC. A brief history of hemoglobins: plant, animal, protist, and bacteria. Proc Natl Acad Sci U S A 1996;93(12):5675–9.
[3] Freitas TAK, Hou S, Dioum EM, Saito JA, Newhouse J, Gonzalez G, et al. Ancestral hemoglobins in Archaea. Proc Natl Acad Sci U S A 2004;101 (17):6675–80.
[4] Dickerson RE, Geis I. Hemoglobin: structure, function, evolution, and pathology. Benjamin-Cummings Publishing Company; 1983.
[5] Burmester T, Hankeln T. Function and evolution of vertebrate globins. Acta Physiol (Oxf) 2014;211(3):501–14.
[6] Pauling L. The Oxygen Equilibrium of Hemoglobin and Its Structural Interpretation. Proc Natl Acad Sci U S A 1935;21(4):186–91.
[7] Pauling L, Coryell CD. The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin. Proc Natl Acad Sci U S A 1936;22(4):210–6.
[8] Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 1951;37(4):205–11.
[9] Pauling L, Corey RB. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. Proc Natl Acad Sci U S A 1951;37(11):729–40.
[10] Pauling L, Itano HA, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. Science 1949;110(2865):543–8.
[11] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 1958;181(4610):662–6.
[12] Anfinsen, C. B., Anfinsen, C. B. The molecular basis of evolution; 1959.
[13] Zuckerkandl E. The evolution of hemoglobin. Sci Am 1965;212(5):110–8.
[14] Fitch WM. An improved method of testing for evolutionary homology. J Mol Biol 1966;16(1):9–16.
[15] Perutz, D. Structure of haemoglobin. Science is Not a Quiet Life: Unravelling the Atomic Mechanism of Haemoglobin **4**; 1997. p. 169.
[16] Lesk A. Introduction to protein science: architecture, function, and genomics. Oxford University Press; 2010. p. 285–346.
[17] Weber RE, Vinogradov SN. Nonvertebrate hemoglobins: functions and molecular adaptations. Physiol Rev 2001;81(2):569–628.
[18] Vinogradov SN, Tinajero-Trejo M, Poole RK, Hoogewijs D. Bacterial and archaeal globins - a revised perspective. Biochim Biophys Acta 2013;1834 (9):1789–800.
[19] Appleby CA. Leghemoglobin and Rhizobium respiration. Annu Rev Plant Physiol 1984;35(1):443–78.
[20] Appleby CA, Tjepkema JD, Trinick MJ. Hemoglobin in a nonleguminous plant, parasponia: possible genetic origin and function in nitrogen fixation. Science 1983;220(4600):951–3.
[21] Bogusz D, Appleby CA, Landsmann Jörg, Dennis ES, Trinick MJ, Peacock WJ. Functioning haemoglobin genes in non-nodulating plants. Nature 1988;331 (6152):178–80.
[22] Hill R, Hargrove M, Arredondo-Peter R. Phytoglobin: a novel nomenclature for plant globins accepted by the globin community at the 2014 XVIII conference on Oxygen-Binding and Sensing. Proteins 2016;5:212.
[23] Gupta KJ, Hebelstrup KH, Mur LA, Igamberdiev AU. Plant hemoglobins: important players at the crossroads between oxygen and nitric oxide. FEBS Lett 2011;585:3843–9.
[24] Gupta KJ, Igamberdiev AU. The anoxic plant mitochondrion as a nitrite: NO reductase. Mitochondrion 2011;11(4):537–43.
[25] Vigeolas H, Hühn D, Geigenberger P. Nonsymbiotic hemoglobin-2 leads to an elevated energy state and to a combined increase in polyunsaturated fatty acids and total oil content when overexpressed in developing seeds of transgenic Arabidopsis plants. Plant Physiol 2011;155(3):1435–44.
[26] Spyrakis F, Bruno S, Bidon-Chanal A, Luque FJ, Abbruzzetti S, Viappiani C, et al. Oxygen binding to Arabidopsis thaliana AHb2 nonsymbiotic hemoglobin: evidence for a role in oxygen transport. IUBMB Life 2011;63:355–62.

[27] Hou S, Larsen RW, Boudko D, Riley CW, Karatan E, Zimmer M, et al. Myoglobin-like aerotaxis transducers in Archaea and Bacteria. Nature 2000;403(6769):540–4.

[28] Chan MK. Recent advances in heme-protein sensors. Curr Opin Chem Biol 2001;5(2):216–22.

[29] Hou S, Freitas T, Larsen RW, Piatibratov M, Sivozhelezov V, Yamamoto A, et al. Globin-coupled sensors: a class of heme-containing sensors in Archaea and Bacteria. Proc Natl Acad Sci U S A 2001;98(16):9353–8.

[30] Wittenberg JB, Bolognesi M, Wittenberg BA, Guertin M. Truncated hemoglobins: a new family of hemoglobins widely distributed in bacteria, unicellular eukaryotes, and plants. J Biol Chem 2002;277(2):871–4.

[31] Vuletich DA, Lecomte JTJ. A phylogenetic and structural analysis of truncated hemoglobins. J Mol Evol 2006;62(2):196–210.

[32] Pesce A, Couture M, Dewilde S, Guertin M, Yamauchi K, Ascenzi P, et al. A novel two-over-two alpha-helical sandwich fold is characteristic of the truncated hemoglobin family. EMBO J 2000;19:2424–34.

[33] Burmester T, Weich B, Reinhardt S, Hankeln T. A vertebrate globin expressed in the brain. Nature 2000;407(6803):520–3.

[34] Kawada N, Kristensen DB, Asahina K, Nakatani K, Minamiyama Y, Seki S, et al. Characterization of a stellate cell activation-associated protein (STAP) with peroxidase activity found in rat hepatic stellate cells. J Biol Chem 2001;276 (27):25318–23.

[35] Burmester, T., Ebner, B., Weich, B., and Hankeln, T. Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. *Mol Biol Evol* **19**; 2002. p. 416-421.

[36] Trent JT, Hargrove MS. A ubiquitously expressed human hexacoordinate hemoglobin. J Biol Chem 2002;277(22):19538–45.

[37] Roesner, A., Fuchs, C., Hankeln, T., Burmester, T. A globin gene of ancient evolutionary origin in lower vertebrates: evidence for two distinct globin families in animals. *Mol Biol Evol*; 2005. 22, p. 12-20.

[38] Kugelstadt D, Haberkamp M, Hankeln T, Burmester T. Neuroglobin, cytoglobin, and a novel, eye-specific globin from chicken. Biochem Biophys Res Commun 2004;325(3):719–25.

[39] Fuchs C, Burmester T, Hankeln T. The amphibian globin gene repertoire as revealed by the Xenopus genome. Cytogenet Genome Res 2006;112:296–306.

[40] Hoogewijs D, Ebner B, Germani F, Hoffmann FG, Fabrizius A, Moens L, et al. Androglobin: a chimeric globin in metazoans that is preferentially expressed in Mammalian testes. Mol Biol Evol 2012;29(4):1105–14.

[41] Isogai Y, Ishida M. Design of a novel heme protein with a non-heme globin scaffold. Biochemistry 2009;48(34):8136–42.

[42] Pastore A, Lesk AM. Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. Proteins 1990;8 (2):133–55.

[43] Murray JW, Delumeau O, Lewis RJ. Structure of a nonheme globin in environmental stress signaling. Proc Natl Acad Sci U S A 2005;102 (48):17320–5.

[44] Stranzl GR, Santelli E, Bankston LA, La Clair C, Bobkov A, Schwarzenbacher R, et al. Structural insights into inhibition of Bacillus anthracis sporulation by a novel class of non-heme globin sensor domains. J Biol Chem 2011;286 (10):8448–58.

[45] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596 (7873):583–9.

[46] Letunic, I., Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* **46**;2018. p. D493-D496.

[47] Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol Direct 2012;7:18.

[48] Zhang D, Iyer LM, Burroughs AM, Aravind L. Resilience of biochemical activity in protein domains in the face of structural divergence. Curr Opin Struct Biol 2014;26:92–103.

[49] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[50] Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform 2009;23:205–11.

[51] Hardison, R. C. Evolution of hemoglobin and its genes. *Cold Spring Harb Perspect Med* **2**;2012. a011627.

[52] Ermler U, Siddiqui RA, Cramm R, Friedrich B. Crystal structure of the flavohemoglobin from Alcaligenes eutrophus at 1.75 A resolution. EMBO J 1995;14(24):6067–77.

[53] Tarricone C, Galizzi A, Coda A, Ascenzi P, Bolognesi M. Unusual structure of the oxygen-binding site in the dimeric bacterial hemoglobin from Vitreoscilla sp. Structure 1997;5(4):497–507.

[54] Freitas TA, Hou S, Alam M. The diversity of globin-coupled sensors. FEBS Lett 2003;552:99–104.

[55] Hill DR, Belbin TJ, Thorsteinsson MV, Bassam D, Brass S, Ernst A, et al. GlbN (cyanoglobin) is a peripheral membrane protein that is restricted to certain Nostoc spp. J Bacteriol 1996;178(22):6587–98.

[56] Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33: W244–8.

[57] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software: Pract Exp 1991;21(11):1129–64.

[58] Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 2004;20(18):3702–4.

[59] Pei, J., Grishin, N. V. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23**; 2007. p. 802-808.

[60] Lassmann T, Sonnhammer EL. Kalign–an accurate and fast multiple sequence alignment algorithm. BMC Bioinf 2005;6:298.

[61] Kumar, S., Stecher, G., Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**; 2016. p. 1870-1874.

[62] Tan Y, Schneider T, Shukla PK, Chandrasekharan MB, Aravind L, Zhang D. Unification and extensive diversification of M/Orf3-related ion channel proteins in coronaviruses and other nidoviruses. Virus Evol 2021;7:veab014.

[63] Holm, L. Benchmarking fold detection by DaliLite v.5. *Bioinformatics* **35**;2019. p. 5326-5327.

[64] DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* **40**; 2002. p. 82-92.

[65] Aravind L. Guilt by association: contextual information in genome analysis. Genome Res 2000;10:1074–7.

[66] Shmakov SA, Faure G, Makarova KS, Wolf YI, Severinov KV, Koonin EV. Systematic prediction of functionally linked genes in bacterial and archaeal genomes. Nat Protoc 2019;14(10):3013–31.

[67] Gantt E. Phycobilisomes. *Annual Review of*. Plant Physiol 1981;32(1):327–47.

[68] MacColl R. Cyanobacterial phycobilisomes. J Struct Biol 1998;124(2-3):311–34.

[69] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.

[70] Teh, A. H., Saito, J. A., Baharuddin, A., Tuckerman, J. R., Newhouse, J. S., Kanbe, M., Newhouse, E. I., Rahim, R. A., Favier, F., Didierjean, C., Sousa, E. H., Stott, M. B., Dunfield, P. F., Gonzalez, G., Gilles-Gonzalez, M. A., Najimudin, N., and Alam, M. Hell's Gate globin I: an acid and thermostable bacterial hemoglobin resembling mammalian neuroglobin. *FEBS Lett* **585**;2011. p. 3250-3258.

[71] Nardini M, Pesce A, Milani M, Bolognesi M. Protein fold and structure in the truncated (2/2) globin family. Gene 2007;398(1-2):2–11.

[72] Schirmer T, Huber R, Schneider M, Bode W, Miller M, Hackert ML. Crystal structure analysis and refinement at 2.5 A of hexameric C-phycocyanin from the cyanobacterium Agmenellum quadruplicatum. The molecular model and its implications for light-harvesting. J Mol Biol 1986;188:651–76.

[73] Dixon MM, Huang S, Matthews RG, Ludwig M. The structure of the C-terminal domain of methionine synthase: presenting S-adenosylmethionine for reductive methylation of B12. Structure 1996;4(11):1263–75.

[74] Padmanabhan S, Jost M, Drennan CL, Elías-Arnanz M. A New Facet of Vitamin B12: Gene Regulation by Cobalamin-Based Photoreceptors. Annu Rev Biochem 2017;86(1):485–514.

[75] Jacob-Dubuisson F, Mechaly A, Betton J-M, Antoine R. Structural insights into the signalling mechanisms of two-component systems. Nat Rev Microbiol 2018;16(10):585–93.

[76] Ulrich LE, Koonin EV, Zhulin IB. One-component systems dominate signal transduction in prokaryotes. Trends Microbiol 2005;13(2):52–6.

[77] Casino P, Miguel-Romero L, Marina A. Visualizing autophosphorylation in histidine kinases. Nat Commun 2014;5:3258.

[78] Bourret RB. Receiver domain structure and function in response regulator proteins. Curr Opin Microbiol 2010;13(2):142–9.

[79] Krell T, Lacal J, Busch A, Silva-Jiménez H, Guazzaroni M-E, Ramos JL. Bacterial sensor kinases: diversity in the recognition of environmental signals. Annu Rev Microbiol 2010;64(1):539–59.

[80] Cheung J, Hendrickson WA. Sensor domains of two-component regulatory systems. Curr Opin Microbiol 2010;13(2):116–23.

[81] Upadhyay AA, Fleetwood AD, Adebali O, Finn RD, Zhulin IB, Schlessinger A. Cache Domains That are Homologous to, but Different from PAS Domains Comprise the Largest Superfamily of Extracellular Sensors in Prokaryotes. PLoS Comput Biol 2016;12(4):e1004862.

[82] Jung K, Fried L, Behr S, Heermann R. Histidine kinases and response regulators in networks. Curr Opin Microbiol 2012;15(2):118–24.

[83] Aravind, L., and Ponting, C. P. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. *FEMS Microbiol Lett* **176**; 1999. p. 111-116.

[84] Watts KJ, Taylor BL, Johnson MS. PAS/poly-HAMP signalling in Aer-2, a soluble haem-based sensor. Mol Microbiol 2011;79:686–99.

[85] Stewart V, Chen L-L. The S helix mediates signal transmission as a HAMP domain coiled-coil extension in the NarX nitrate sensor from Escherichia coli K-12. J Bacteriol 2010;192(3):734–45.

[86] Sharma AK, Rigby AC, Alper SL. STAS domain structure and function. Cell Physiol Biochem 2011;28(3):407–22.

[87] Aravind L, Koonin EV. The STAS domain - a link between anion transporters and antisigma-factor antagonists. Curr Biol 2000;10(2):R53–5.

[88] Aravind L, Anantharaman V, Balaji S, Babu M, Iyer L. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 2005;29(2):231–62.

[89] Wan T, Li S, Beltran DG, Schacht A, Zhang L, Becker DF, et al. Structural basis of non-canonical transcriptional regulation by the sigmaA-bound iron-sulfur protein WhiB1 in M. tuberculosis. Nucleic Acids Res 2020;48:501–16.

[90] Chou S-H, Galperin MY, O'Toole GA. Diversity of Cyclic Di-GMP-Binding Proteins and Mechanisms. J Bacteriol 2016;198(1):32–46.

[91] Zschiedrich CP, Keidel V, Szurmant H. Molecular Mechanisms of Two-Component Signal Transduction. J Mol Biol 2016;428(19):3752–75.

[92] Galperin MY, Natale DA, Aravind L, Koonin EV. A specialized version of the HD hydrolase domain implicated in signal transduction. J Mol Microbiol Biotechnol 1999;1:303–5.

[93] Marles-Wright J, Lewis RJ. Stress responses of bacteria. Curr Opin Struct Biol 2007;17(6):755–60.

[94] Iber D, Clarkson J, Yudkin MD, Campbell ID. The mechanism of cell differentiation in Bacillus subtilis. Nature 2006;441(7091):371–4.

[95] Arigoni F, Guerout-Fleury A-M, Barak I, Stragier P. The SpoIIE phosphatase, the sporulation septum and the establishment of forespore-specific transcription in Bacillus subtilis: a reassessment. Mol Microbiol 1999;31(5):1407–15.

[96] Graham-Lorence S, Peterson JA, Amarneh B, Simpson ER, White RE. A three-dimensional model of aromatase cytochrome P450. Protein Sci 1995;4(6):1065–80.

[97] Sciara G, Kendrew SG, Miele AE, Marsh NG, Federici L, Malatesta F, et al. The structure of ActVA-Orf6, a novel type of monooxygenase involved in actinorhodin biosynthesis. EMBO J 2003;22:205–15.

[98] Tinikul R, Chunthaboon P, Phonbuppha J, Paladkong T. Bacterial luciferase: Molecular mechanisms and applications. Enzymes 2020;47:427–55.

[99] Francis VI, Porter SL. Multikinase Networks: Two-Component Signaling Networks Integrating Multiple Stimuli. Annu Rev Microbiol 2019;73(1):199–223.

[100] Henry JT, Crosson S. Ligand-binding PAS domains in a genomic, cellular, and structural context. Annu Rev Microbiol 2011;65(1):261–86.

[101] Taylor BL, Zhulin IB. PAS domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev 1999;63(2):479–506.

[102] Yang X, Stojković EA, Ozarowski WB, Kuk J, Davydova E, Moffat K. Light Signaling Mechanism of Two Tandem Bacteriophytochromes. Structure 2015;23(7):1179–89.

[103] Podust LM, Ioanoviciu A, Ortiz de Montellano PR. 2.3 A X-ray structure of the heme-bound GAF domain of sensory histidine kinase DosT of Mycobacterium tuberculosis. Biochemistry 2008;47:12523–31.

[104] Fushimi K, Narikawa R. Cyanobacteriochromes: photoreceptors covering the entire UV-to-visible spectrum. Curr Opin Struct Biol 2019;57:39–46.

[105] Glazer AN. Light guides. Directional energy transfer in a photosynthetic antenna. J Biol Chem 1989;264(1):1–4.

[106] Puzorjov, A., McCormick, A. J. Phycobiliproteins from extreme environments and their potential applications. *J Exp Bot* **71**;2020. p. 3827-3842.