Sequence conservation, domain architectures, and phylogenetic distribution of the HD-GYP type c-di-GMP phosphodiesterases

Michael Y. Galperin^{1*} and Shan-Ho Chou^{2,3}

¹ National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

² State Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, P. R. China

³ Institute of Biochemistry and Agricultural Biotechnology Center, National Chung Hsing University, Taichung 40227, Taiwan, Republic of China

*For correspondence. E-mail: galperin@ncbi.nlm.nih.gov, Tel: +1-301-435-5910

ORCID numbers

M. Y. Galperin, [0000-0002-2265-5572](https://orcid.org/0000-0002-2265-5572)

S.-H. Chou: [0000-0002-5202-8238](https://orcid.org/0000-0002-5202-8238)

Running title: Sequence analysis of the HD-GYP domain

ABSTRACT

The HD-GYP domain, named after two of its conserved sequence motifs, was first described in 1999 as a specialized version of the widespread HD phosphohydrolase domain that had additional highly conserved amino acid residues. Domain associations of HD-GYP indicated its involvement in bacterial signal transduction and distribution patterns of this domain suggested that it could serve as a hydrolase of the bacterial second messenger c-di-GMP, in addition to or instead of the EAL domain. Subsequent studies confirmed the ability of various HD-GYP domains to hydrolyze c-di-GMP to linear pGpG and/or GMP. Certain HD-GYP-containing proteins hydrolyze another second messenger, cGAMP, and some HD-GYP domains participate in regulatory protein-protein interactions. The recently solved structures of HD-GYP domains from four distinct organisms clarified the mechanisms of c-di-GMP binding and metal-assisted hydrolysis. However, the HD-GYP domain is poorly represented in public domain databases, which causes certain confusion about its phylogenetic distribution, functions, and domain architectures. Here, we present a refined sequence model for the HD-GYP domain and describe the roles of its most conserved residues in metal and/or substrate binding. We also calculate the numbers of HD-GYPs encoded in various genomes and list the most common domain combinations involving HD-GYP, such as the RpfG (REC-HD-GYP), Bd1817 (DUF3391-HD-GYP), and PmGH (GAF-HD-GYP) protein families. We also provide the descriptions of six HD-GYP-associated domains, including four novel integral membrane sensor domains. This work is expected to stimulate studies of diverse HD-GYP-containing proteins, their N-terminal sensor domains and the signals to which they respond.

IMPORTANCE

The HD-GYP domain forms class II of c-di-GMP phosphodiesterases that control the cellular levels of the universal bacterial second messenger c-di-GMP and therefore affect flagellar and/or twitching motility, cell development, biofilm formation, and, often, virulence. Despite more than 20 years of research, HD-GYP domains are insufficiently characterized; they are often confused with ‘classical’ HD domains that are involved in various housekeeping activities and may participate in signaling, hydrolyzing (p)ppGpp and c-di-AMP. This work provides an updated description of the HD-GYP domain, including its sequence conservation, phylogenetic distribution, domain architectures, and the most widespread HD-GYP-containing protein families. This work shows that HD-GYP domains are widespread in many environmental bacteria and are predominant c-di-GMP hydrolases in many lineages, including clostridia and deltaproteobacteria.

Keywords: bacteria, virulence, biofilm, c-di-GMP, c-di-AMP, signal transduction, phosphodiesterase

INTRODUCTION

Bis-(3',5')-cyclic dimeric guanosine monophosphate (c-di-GMP) was originally discovered as the activator of bacterial cellulose synthase (1) and was subsequently recognized as a universal bacterial second messenger (2-6). In many bacteria, c-di-GMP controls a variety of regulatory processes, including protein and polysaccharide secretion, motility, cell development, and biofilm formation (7-9). In some important bacterial pathogens, c-di-GMP is involved in the regulation of virulence (7-12). The cellular levels of c-di-GMP, which typically range from 40 to 80 nM (13), are tightly controlled by multiple synthases (diguanylate cyclases), containing the GGDEF domain, and c-di-GMP hydrolases (phosphodiesterases) that are found in two structurally distinct forms of the EAL and HD-GYP domains. All these domains are widespread in bacteria and are often encoded in multiple copies in the same genome. These domains can also be found in catalytically inactive forms which often bind c-di-GMP and/or participate in protein-protein interactions. Some c-di-GMP molecules appear to form an intracellular pool of this compound whereas others function locally without joining that pool (13-15). The structure and catalytic mechanisms of c-di-GMP synthesis by the GGDEF domains and its hydrolysis by the EAL domains have been characterized in much detail (7, 16-21).

The second c-di-GMP-hydrolyzing domain, HD-GYP, has been the subject of far fewer studies and very few comprehensive reviews (22-25). This domain, named after two of its conserved sequence motifs, was initially described in 1999 (3) as a specialized version of the widespread HD phosphohydrolase domain, which contains the metal-binding His-Asp active-site motif (26). In addition to the wide variety of phosphohydrolases (phosphatases, phosphodiesterases, nucleotidases, nucleases, nucleotidyltransferases), including the Cas3 nuclease of the CRISPR-Cas system (27, 28), the HD domain superfamily includes certain hydratases, oxygenases, and lyases (29, 30). Compared to the consensus ~100-aa HD domain (26), the HD-GYP domain was defined as a longer (166-aa) unit with a large number of additional conserved residues including the characteristic GYP sequence motif (3).

Domain architectures of HD-GYP-containing proteins included its combinations with periplasmic ligand-binding sensor domains, the phosphoacceptor (receiver, REC) domain of the two-component signal transduction system, and the GGDEF domain that was previously identified in diguanylate cyclases and c-di-GMP phosphodiesterases (2). These domain associations indicated the involvement of HD-GYP in bacterial signal transduction (3). Based on the numbers of GGDEF, EAL, and HD-GYP domains encoded in various bacterial genomes, the latter two domains were predicted to perform the same function, namely, serve as diguanylate phosphodiesterases, catalyzing the hydrolysis of c-di-GMP molecules synthesized by the GGDEF domain (3). Two years later, in 2001, a partial sequence logo of the HD-GYP domain was published, highlighting the conservation of the predicted metal-binding residues (4).

Subsequent studies of various proteins containing the HD-GYP domain confirmed its ability to hydrolyze c-di-GMP, either introducing a single nick resulting in the production of a linear 5'-phosphoguanylyl-(3',5')-guanosine (pGpG) or hydrolyzing it further to two molecules of GMP (31-40). Some HD-GYP domains also hydrolyzed its sibling 3',3'-cyclic-GMP-AMP (cGAMP) (41-43) but not c-di-AMP or any other cyclic dinucleotide second messengers (41). In the past several years, crystal structures of HD-GYP domains from four different organisms have been solved (42, 44-46), shedding

light on the mechanisms of metal-assisted hydrolysis of c-di-GMP and cGAMP. Like the ‘classical’ HD domain and in agreement with the original predictions (3, 26), HD-GYP is an all- α domain that consists of at least 7 α -helices.

Despite this progress, HD-GYP domain has been far less characterized than GGDEF and EAL domains and was even called their “neglected small sibling” (22). One of the reasons for that is that, unlike GGDEF and EAL domains, HD-GYP domain is not encoded in *Escherichia coli*, *Salmonella enterica*, *Bacillus subtilis*, and some other model organisms (3, 4), see https://www.ncbi.nlm.nih.gov/Complete_Genomes/c-di-GMP.html. Another reason is the absence of an appropriate sequence model of the HD-GYP domain in most public domain databases. As a result, HD-GYP domains are often confused with the ‘classical’ HD domains, which hampers the analysis of their domain architectures and obscures their specific functions. This work aims at filling that void by providing a detailed analysis of sequence conservation, phylogenetic distribution, domain architectures, and the c-di-GMP binding mechanisms of the HD-GYP domain.

RESULTS AND DISCUSSION

Sequence model and structure of the HD-GYP domain

The original description of the HD-GYP domain (3) included a 166-aa alignment that, in addition to the HD motif, included many additional conserved residues including the conserved HHE_xh(D/N)G_xGYP sequence motif, where *h* indicates a hydrophobic acid residue and *x* – any residue (3). This alignment was subsequently used as the basis for the sequence logo representations of the HD-GYP domain in the reviews of c-di-GMP metabolism [Fig. 1c in reference (4) and Fig. 3c in reference (7)]. Similar alignments appeared in the papers describing the crystal structures of HD-GYP domains from *Bdellovibrio bacteriovorus*, *Persephonella marina*, *Pseudomonas aeruginosa*, and *Vibrio cholerae* (42, 44-46). **Figure 1** presents an expanded structure-based sequence alignment that captures the key conserved elements of the HD-GYP domain. This alignment features an unusually high number of conserved residues; functional roles of some of them, deduced from HD-GYP domain structures (42, 44-46), are listed in **Table 1**. This 180-aa alignment is longer than most previous ones and includes three C-terminal α -helices, which, despite their common presence in HD-GYP domains, show little sequence conservation (Figure 1). In multi-domain proteins, the C-terminal HD-GYP domain is often preceded by the S-helix (47); the N-terminal HD-GYP domain of the *Aquifex aeolicus* protein Aq_2027 (GenBank accession number [AAC07788.1](https://www.ncbi.nlm.nih.gov/nuclot/AAC07788.1)), which was used in the initial description of the HD-GYP domain (3), is also preceded by an α -helix (not shown in Figure 1).

Unfortunately, most current protein domain databases lack a proper sequence model of the HD-GYP domain. For example, the Pfam database (48) contains two entries for HD-GYP-related domains, HD ([PF01966](https://pfam.xfam.org/family/PF01966)) and HD_5 ([PF13487](https://pfam.xfam.org/family/PF13487)), neither of which covers the entire sequence of HD-GYP. As shown in Figure 1, the 113-aa PF01966 sequence model misses the N-terminal loop that often carries a metal-binding acidic residue (Glu or Asp) and three C-terminal α -helices of HD-GYP. Some sequences in PF01966 contain the GYP motif and likely represent HD-GYP domains, while others do not. Judging from the conservation of the key active-site residues (Figure 1), full-fledged HD-GYP domains comprise at least 8,500 out of nearly 50,000 members of Pfam’s “full” list of HD ([PF01966](https://pfam.xfam.org/family/PF01966)) domains. The 64-aa PF13487 sequence model misses the eponymous HD sequence motif and three other metal-binding

residues. Still, an analysis of full-length sequences of ca. 7,700 members of Pfam's "full" list of HD_5 (PF13487) domains shows that at least 80% of them represent HD-GYP domains. Pfam domain entries HD_2 (PF_12917), HD_3 (PF13023), HD_4 (PF13328), and HD_6 (PF18019) do not seem to include any HD-GYP domains.

Other protein domain databases also have problems in recognizing the HD-GYP domain. The 124-aa HDc sequence model SM00471 in the SMART database (49) and its derivatives HDc (cd00077) in the Conserved Domain Database (CDD)(50) and IPR003607 in the InterPro (51) cover the metal-binding residues but lack the C-terminal α -helices. The cd00077 profile also misses the conserved Lys residue (Lys235 in the 4MCW structure, Figure 1) that is involved in c-di-GMP binding. By contrast, the two entries in the COG database (52), COG3437 and COG2206, cover the entire HD-GYP domain but contain additional N-terminal sequences from the receiver (REC, PF00072 in Pfam) domain in the former and the DUF3391 (PF11871) domain in the latter (see below for details). As a result, the HD-GYP entry PS51832 in the ProSite database (53) and its InterPro derivative IPR037522 are the only ones that include the entire HD-GYP domain without any extra amino acid residues. As of October 1st, 2021, this InterPro entry IPR037522 contained ~80,000 entries including 9 crystal structures of the HD-GYP domain from the Protein DataBank (PDB). The structure-based sequence alignment shown in Figure 1 and its expanded version shown in **Figure S1** in the Supplemental file 1 should allow devising suitable sequence models for distinguishing true HD-GYPs from other HD superfamily domains, such as HDc (26) or HDOD (54, 55). As a first step towards this goal, Supplemental file 2 provides a suitable query sequence of the HD-GYP domain and an associated position-specific scoring matrix (PSSM) for use in PSI-BLAST (56) searches and a hidden Markov model (HMM) for use in HMMer (57) searches of protein sequence databases.

Substrate binding and hydrolysis by HD-GYP domains

Studies of the c-di-GMP- and cGAMP-specific phosphodiesterase activities of the HD-GYP-containing proteins (listed in **Table S1** in the supplemental materials) revealed the key role in catalysis of two metal atoms, usually Fe²⁺ or Mn²⁺, and the metal-binding residues, highlighted in Figure 1 and listed in Table 1 (41, 45). These include the partially conserved Glu residue in the N-terminal loop (E185 in the structure of *P. marina* PmGH, PDB ID: 4MCW, Figure 1). This acidic residue (Glu or Asp) participates in binding the third metal atom; phylogenetic analyses showed that, depending on the presence or absence of this Glu residue, HD-GYP domains could be split into two families carrying, respectively, trinuclear or binuclear metal centers (45, 46, 58). For their c-di-GMP phosphodiesterase activity, experimentally studied HD-GYP domains typically require two ferrous (Fe²⁺) ions, oxidation of either one of them inactivates the enzyme (36). The phosphodiesterase activity has also been observed with manganese (Mn²⁺) ions (36, 39). Some HD-GYP domains could bind Co²⁺, Ni²⁺, Mg²⁺, Zn²⁺, Ca²⁺, and Cu²⁺ with similar affinities (K_d around 1 μ M), and some of these metal complexes showed comparable catalytic activity (39, 42, 46, 58), reviewed in (25).

The presence of the E185 residue and, accordingly, of the trinuclear metal center, is an important determinant of the ability of the HD-GYP domains to catalyze the second reaction, hydrolysis of pGpG to GMP. In several independent studies, E185-carrying HD-GYP domains of such proteins as VC1295 and VCA0210 from *V. cholerae*, PmGH from *P. marina*, TM0186 from *T. maritima*, PA4108 from *P.*

aeruginosa, and DVUA0086 from *Desulfovibrio vulgaris* were shown to hydrolyze c-di-GMP all the way to GMP (36-39, 45), whereas closely related enzymes missing that residue (e.g. containing a Tyr at that position) were unable to do so. However, this requirement is not absolute: PA4781 and VCA0681 proteins, which lack Glu or Asp at that position (Figure 1), have been seen producing GMP as well (35, 37, 38). On the other hand, the isolated HD-GYP domain of the *D. vulgaris* protein DVU0722, which has this Glu residue (Figure S1), was reported to hydrolyze c-di-GMP mostly to pGpG and produce very little GMP, at least in the presence of Mn^{2+} (39). Apparently, there are multiple ways of arranging the HD-GYP domain active site to hydrolyze pGpG and there are additional controlling mechanisms.

As for the substrate, the PmGH–c-di-GMP complex [PDB ID: 4MDZ, (45)] shows a c-di-GMP monomer in a V-shape conformation (59). Its binding involves several conserved amino acid residues that are listed in Table 1 and highlighted in purple in Figure 1. Most of these residues interact with just one of the guanine bases (Gua1 in Figure 2), while the second one, Gua2, is primarily bound by main chain carbonyl oxygens of Lys and Gly residues of the highly conserved KPG motif (Lys235 and Gly237 in the PmGH structure, Figures 1 and S1). This organization of the active site accounts for greater flexibility of the Gua2 base and may tolerate its replacement by adenine, which would make the HD-GYP domain able to hydrolyze cGAMP in addition to or instead of c-di-GMP (41-43). The cGAMP hydrolysis could also result from mutations in the Gua1-interacting Arg314-Lys317 pair that forms highly conserved RxxK (or, more precisely, RxY[K/R]) motif (Figure S1) and is largely responsible for the HD-GYP's specificity for c-di-GMP. In several proteins, including Bd2325, Gmet_3476, and Moth_2495 (Figure S1), replacement of Lys317 with Gln or Asn (resulting, respectively, in the RxxQ or RxxN motifs) shifted substrate specificity towards cGAMP (43). However, HdgA (Vnz_24090, GenBank accession number [APE23802.1](https://www.ncbi.nlm.nih.gov/nuclot/111111111)) protein from *Streptomyces venezuelae* (60), in which the Lys residue of the RxxK motif was replaced by Ser (Figure S1) was still able to hydrolyze c-di-GMP (Dr. Natalia Tschowri, personal communication). In some deltaproteobacterial proteins, this motif is RxxH, with Lys residue replaced by His.

The oxygen atoms of the central 12-member ribose-phosphate ring of c-di-GMP (and cGAMP) are bound by Me^{2+} ions that are coordinated by the metal-binding residues of HD-GYP. The Tyr285 residue of the eponymous GYP motif enhances the binding strength and substrate specificity of the HD-GYP domain by forming a hydrogen bond with one of the non-bridging oxygen atoms of the ribose-phosphate ring (Figure 2).

The catalytic mechanism of c-di-GMP hydrolysis by HD-GYP domains is believed to involve a nucleophilic attack on the phosphorus atom of the central ribose-phosphate ring by a hydroxide ion, activated by one or two Fe^{2+} (or other Me^{2+}) ions of the active site (36). The mechanistic aspects of the phosphodiesterase activity of the HD-GYP domain and its allosteric regulation are described in detail in recent reviews (24, 25).

Phylogenetic distribution and abundance of the HD-GYP domain

The HD-GYP domain is often assumed to be far less widespread than the isofunctional EAL domain (24). However, this impression is largely due to the *E. coli*-centric view of the bacterial world. Indeed, while the HD-GYP domain is not encoded in the genomes of *E. coli* or *B. subtilis* (3, 4), it is abundant in

many environmental bacteria. In the current release of the COG database, which aims at covering microbial diversity (52) and includes complete genomes of 1,187 bacterial species from 1,116 genera (typically, a single representative genome per bacterial genus), HD-GYP domains are found in 651 bacterial genomes, compared to 754 that encode EAL domains (**Table S2** in the supplemental materials). Close to a half of bacteria in the list, 586 of 1,187, encode more EAL domains than HD-GYP domains but 252 genomes encode as many or more HD-GYPs than EALs, and 349 do not encode either of them.

Among the organisms covered by COGs, the number of the HD-GYP domains per genome varies from zero to 26 and does not show a clear correlation with the genome size (**Figure 3**). HD-GYP domains show prevalence among certain representatives of the phyla *Chloroflexi*, *Dictyoglomi*, *Planctomycetes*, *Spirochaetes*, *Synergistetes*, and *Thermotogae*, as well as in *Clostridia* and the proteobacterial classes *Deltaproteobacteria*, *Oligoflexia*, and *Zetaproteobacteria* (Table S2). **Table 2** lists some bacterial species with a substantial prevalence of HD-GYP domains over their EAL counterparts. Some of the organisms with multiple HD-GYP domains are extremophiles: thermophiles, halophiles, and alkalophiles, which suggests certain robustness of their ability to hydrolyze c-di-GMP. This trend is seen, for example, in two extremely haloalkaliphilic clostridia, *Halanaerobium hydrogeniformans*, isolated from Soap Lake, WA, and *Natronaerobius thermophilus*, isolated from Wadi An Natrun lake in Egypt. The genome of the former encodes 22 HD-GYP domains and just two EAL domains, the latter encodes 15 HD-GYPs and no EALs (Table 2). As noted previously (3), the syphilis-causing spirochete *Treponema pallidum* encodes three HD-GYP domains and no EALs. Genomes of several apparently benign gut symbionts such as *Adlercreutzia equolifaciens*, *Lawsonia intracellularis*, *Sutterella megalosphaeroides*, and *Veillonella parvula*, encode single HD-GYP-containing enzymes as their only c-di-GMP-hydrolyzing phosphodiesterases. Also interesting is the presence of HD-GYP domains, but not EAL domains, in the ancestral cyanobacteria-like organism *Candidatus Melainabacteria* bacterium MEL.A1 and the early-branching cyanobacterium *Gloeobacter violaceus* PCC7421.

As c-di-GMP signaling machinery appears limited to bacteria, HD-GYP domains, like GGDEF and EAL domains, were previously reported to be absent in archaea and eukaryotes (7). The only reported exception was a single ORF (RCIX2274, GenBank accession number [CAJ37382](#)) from the methanogen *Methanocella arvoryzae* MRE50 that combines REC, PAS, PocR, GGDEF, and HD-GYP domains. The current study found in GenBank 17 eukaryotic (**Table S3**) and 70 archaeal ORFs (**Table S4**) that contained HD-GYP domains. Three HD-GYP-encoding genes from plants had introns, supporting their likely eukaryotic origin. However, one of these ORFs, AXX17_ATUG04560 from *Arabidopsis thaliana* Landsberg *erecta* (61), was absent from the reference sequence of *A. thaliana*. Most eukaryotic HD-GYP-containing proteins were ≥99% identical to certain bacterial proteins (Table S3). Among archaeal HD-GYPs, only the above-mentioned *M. arvoryzae* ORF came from a complete genome, all others were encoded by metagenome-assembled genomes (MAGs) and represent cases of potential bacterial contamination, as discussed, for example, in reference (62). A case in point is a protein with the HD-GYP domain architecture (GenBank accession number [HGZ60150](#)) from *Fervidicoccus fontis* MAG (63) that is not encoded in the complete genome of *F. fontis* strain Kam940 (64). Future studies will show if any of the proteins listed in Tables S3 and S4 are genuine archaeal or eukaryotic ORFs.

Domain architectures of HD-GYP-containing proteins

The most common domain architecture involving the HD-GYP domain is its combination with the two-component receiver (REC) domain, first described in 1999 (3), and subsequently characterized in the nearly identical RpfG proteins from *Xanthomonas axonopodis*, *Xanthomonas campestris*, and *Xanthomonas oryzae* (31, 65, 66, 67). These proteins were shown to regulate biofilm formation and virulence to citrus, crucifers and rice, respectively. The reported c-di-GMP phosphodiesterase activity was fairly low {Ryan, 2006 #211, 68, 69} (the first of these papers was later retracted), in agreement with the absence of a key His residue in their sequences (Figures 1 and S1). However, a recent report of robust c-di-GMP hydrolysis by a very similar RpfG protein (Figure S1) from *Lysobacter enzymogenes*, another member of *Xantomonadaceae* (70), suggests that this His residue might not be essential after all. The proteins with the REC–HD-GYP domain architecture are referred to as the RpfG family of response regulators (71), see **Table 3**. The experimentally characterized RpfG family members from *Thermotoga maritima*, *V. cholerae*, and *Desulfovibrio vulgaris* contain the full set of active site residues (Figures 1 and S1) and have robust c-di-GMP phosphodiesterase activity (34, 36, 38, 39). Catalytically inactive proteins of the RpfG family participate in regulation through their protein-protein interactions (see below).

A fair number of bacterial response regulators have more complex domain architectures with REC and HD-GYP domains separated by one or more PAS or GAF domains. The only experimentally studied protein with the REC–PAS–HD-GYP domain architecture is DVU2933 from *D. vulgaris*. This protein could not hydrolyze c-di-GMP (at least without phosphorylation of the REC domain) but its isolated HD-GYP domain showed Mn²⁺-dependent catalytic activity (39). The *dvu2933* gene is part of an operon that regulates the *lpxC* gene, which is involved in lipid A biosynthesis, but its specific role is unknown (39). In Table 3, proteins with REC–PAS–HD-GYP domain architecture are referred to as response regulators of the RpfX family, in keeping with the RpfG designation for the REC–HD-GYP domain architecture. The central PAS (or GAF) domain is expected to modulate signal transduction from the phosphoacceptor REC domain to the HD-GYP domain, controlling the c-di-GMP phosphodiesterase activity of the protein.

The HD-GYP domain is also involved in a variety of other domain combinations, see Table 3 for selected examples. As noted above, owing to the absence of the standard sequence model for the HD-GYP domain, distinguishing these domains from ‘classical’ HD domains could be challenging. However, analysis of the domain architectures of these proteins revealed a clear distinction: while ‘classical’ HD domains are typically associated with house-keeping enzymatic activities, HD-GYP domains are usually involved in c-di-GMP-dependent signaling and associated with other signaling domains, either sensory (ligand-binding) or signal transduction domains, such as REC, GAF, PAS, GGDEF, and/or GGDEF-EAL pair. Indeed, a CD-search showed that out of 1,320 response regulators listed in Pfam as combining REC and HD (PF01966) domains and 1,010 sequences with the REC–HD_5 domain combination, at least 90% contain full-length HD-GYP domains, i.e. belong to the RpfG family. With the GAF domain, the picture was the same: all domain combinations listed in Pfam as GAF–HD, GAF_2–HD, GAF_3–HD, GAF_2–GAF–HD, GAF_2–GAF_2–HD, GAF–HD_5, and GAF–HD–HD_5 were found to contain C-terminal HD-GYP domains. Likewise, sequence analysis of the DUF3391 domain (see below) showed that all its combinations with HD and HD_5 domains listed in Pfam represent genuine HD-GYP

domains. This was also true for one-component signaling proteins combining the HD domain with periplasmic sensory domains, initially described in (3). In all 29 cases of the PBPb–HD domain combination, 18 out of 19 cases of CHASE2–HD domain combinations, and all 33 cases of dCache_1–HD, dCache_1–HD_5, and dCache_1–GAF–HD_5 domain combinations, listed in Pfam, the output domain proved to be HD-GYP. Associations with PAS domains were not so clear-cut: all PAS_3–HD, PAS_9–HD, and PAS_9–PAS_3–HD entries contained HD-GYP domains, whereas in PAS–HD and PAS_4–HD combinations, up to a quarter of the sequences represented either ‘classical’ HD domain or an inactivated HD-GYP that has lost most of its metal-binding residues.

Novel HD-GYP-associated domains

A great majority of HD-GYP domain architectures in Pfam (48) show HD or HD_5 as the only recognized domain in the entire protein. The remaining portions of these proteins can be expected to harbor other, highly diverged, poorly characterized, or unrecognized protein domains. Here we provide descriptions of two such domains, which are particularly widespread and are referred to as domains of unknown function, DUF3391 and DUF3369, in Pfam (48). We also describe four novel integral membrane domains that are likely to be sensing diverse ligands to regulate c-di-GMP hydrolysis by downstream HD-GYP domains.

DUF3391. The high-resolution crystal structure of *B. bacteriovorus* Bd1817 (PDB ID: 3TM8 (44)) revealed a 120-aa N-terminal domain, formed by four β -strands with a 3_{10} helix between strands 3 and 4, followed by two long α -helices (**Figure S3**); it is connected to the HD-GYP domain of Bd1817 by a long S-helix. In Pfam, this domain is referred to as Domain of Unknown Function (DUF3391, [PF11871](#)). The DUF3391–HD-GYP domain combination, referred to as the Bd1817 family in Table 3, is widespread in bacteria. Among others, it is found in spirochetes and comprises the only HD-GYP-containing c-di-GMP phosphodiesterase in most *Borellia* spp., including the Lyme disease-causing *B. burgdorferi* (33), and one of three such proteins in *Treponema pallidum*, the causative organism of syphilis (72). It should be noted, however, that while the *Treponema* proteins contain full-size DUF3391 domain, the proteins from *Borellia* spp., including the experimentally characterized BB_0374 (33), are missing the first 40 amino acid residues of DUF3391 (Figure S3A). The DUF3391–HD-GYP domain combination has been seen in two other experimentally characterized proteins, PA4108 from *P. aeruginosa* and VC2497 from *V. cholerae*, both of which hydrolyzed c-di-GMP to pGpG and could also convert pGpG to GMP (37, 38). Further, PA4108 showed catalytic activity when expressed in *E. coli* (37), indicating that the DUF3391 domain did not interfere with the activity of the downstream HD-GYP domain and could promote formation of the active, dimeric form of the enzyme. Many genomes contain multiple proteins of the Bd1817 family, thus, *Shewanella* spp. often contain three or even four such proteins (Figure S3A) (Martin-Rodriguez *et al.*, submitted). An analysis of the domain architectures of DUF3391-containing proteins showed that almost all of them featured the DUF3391–HD-GYP combination.

The 136-aa HMM logo of the DUF3391 domain that is available on the Pfam website <https://pfam.xfam.org/family/PF11871#tabview=tab4> shows only a couple of conserved residues, Gly12 and Trp22. However, this could be attributed to the presence of variable-length Pro-rich inserts between $\alpha 1$ and $\alpha 2$, whose lengths range from zero in Bd1817 to 48 amino acid residues in GNIT_0984 protein from *Glaciecola nitratreducens* (Figure S3, UniProt entry G4QJZ9_GLANF). Collapsing these

low-complexity inserts revealed a conserved core structure with several more conserved residues. Unfortunately, the sequence of Bd1817 is so far from the domain consensus (Figure S3B) that it does not allow making any predictions regarding ligand-binding surface(s) of DUF3391. The structural prediction by trRosetta {Yang, 2020 #287} shows the positions of the most conserved residues, including the hydrophobic stacking of Phe29 with Phe35 and potentially also Phe15 and His27 (Figure S3C). However, the roles of other residues, including the highly conserved Met14, remain obscure. It remains to be seen if DUF3391 binds its ligand(s) with the flat surface of its three-stranded β -sheet or through specific interaction with Met14 and other conserved residues.

DUF3369 (PF11849 in Pfam) is a widespread GAF-like cytoplasmic domain that is frequently associated with the HD-GYP domain and can also be found in diguanylate cyclases, sensor histidine kinases, and hybrid GGDEF–EAL domain proteins (**Figure S4**). Sequence analysis of Pfam entries with DUF3369–HD and DUF3369–HD_5 domain combinations showed that almost all of them contain full-length HD-GYP domains at their C-termini and N-terminal REC domains (sometimes, diverged ones), so they could be considered response regulators of the two-component signal transduction system (54). Two proteins with the REC–DUF3369–HD-GYP domain architecture (the RpfY family in Table 3) have been experimentally characterized. One, VCA0210 from *V. cholerae*, has been mentioned above as a phosphodiesterase capable of hydrolyzing c-di-GMP to GMP (38). The expression of the *vca0210* gene was regulated by quorum sensing and its overexpression resulted in increased motility and decreased biofilm formation (38). In addition to c-di-GMP, VCA0210 was also able to hydrolyze cGAMP to linear dinucleotide (41). Another experimentally studied protein with the same domain architecture, DVU0722 from *D. vulgaris*, did not appear to hydrolyze c-di-GMP, at least without phosphorylation of the REC domain. However, its isolated HD-GYP domain showed Mn^{2+} -dependent catalytic activity, producing pGpG and very little GMP (39). While no crystal structure of DUF3369 has been determined so far, a search with HHpred (73) suggests its structural similarity to GAF. Structural prediction by trRosetta (74), available on the Pfam website shows (48) does not show clustering of the conserved residues of DUF3369 (Figure S4B), so the ligand that it binds, if any, remains obscure. Anyway, like the PAS domain in the RpfX family, the DUF3391 domain in RpfY family proteins likely modulates signal transduction from REC to the HD-GYP domain, controlling its phosphodiesterase activity. Remarkably, as seen in Figure S4A, all four DUF3369-containing proteins of *Desulfomicrobium baculatum* belong to the RpfY family (i.e. have the REC–DUF3369–HD-GYP domain architecture). In contrast, five DUF3369-containing proteins of *Azospirillum lipoferum* are far more diverse: two of them have HD-GYP as the output domain, one has the GGDEF-EAL combination, and two more are histidine kinases with distinct domain architectures (Figure S4A).

MASE9. The original description of the HD-GYP domain (3) identified its fusions with periplasmic sensory domains, subsequently identified as PBPb (TM1170 protein in *T. maritima*) and dCache (TM1682 protein in *T. maritima*) domains. In contrast, in *T. maritima* TM1467 protein, HD-GYP was linked to an N-terminal integral membrane fragment (3) that had not been characterized until now. Sequence analysis of this N-terminal fragment revealed an integral membrane protein domain with 8 predicted transmembrane (TM) segments, two of which were too short to completely traverse the membrane and formed a likely intramembrane hairpin (**Figure S5**). Five positions were occupied by conserved aromatic residues (Phe or Tyr) and 3 positions by hydrophilic (Asn or Ser) residues. By

analogy with the previously defined membrane-associated sensor domains (MASE1 through MASE8 (75-77), Martin-Rodriguez *et al.*, submitted), this domain was named MASE9. In addition to the MASE9–HD-GYP domain combination, MASE9 was also found in several proteins with the MASE9–HD-GYP–GAF–GGDEF domain architecture, as well as in the association with the GGDEF–EAL domain combination (Figure S5), confirming that it represents a promiscuous sensory domain. The poor conservation of the short periplasmic segments, coupled with the presence of several charged residues in the hydrophobic core of the membrane, suggests that ligand binding by MASE9 occurs within the membrane. The only experimentally characterized MASE9-containing protein was HdgA (Vnz_24090) from *Streptomyces venezuelae*, which was poorly expressed and therefore did not appear to have a major role in the c-di-GMP metabolism of this organism (60).

MASE10. The VC1295 protein of *V. cholerae* was initially characterized as one of those whose expression was repressed by bile acids (78). In its sequence, the HD-GYP domain is preceded by a predicted transmembrane domain and a HAMP domain (38). VC1295 has been shown to bind c-di-GMP (79) and hydrolyze it to both pGpG and GMP (38). The VC1295 entries in UniProt, Pfam, and SMART databases all predicted its N-terminal domain to consist of 6 TM segments but provided no information on its membrane topology. Sequence analysis identified similar 6TM N-terminal regions in three classes of signaling proteins: c-di-GMP phosphodiesterases with HD-GYP domains, adenylate/guanylate cyclases producing cAMP or cGMP, and Ser/Thr protein phosphatases of the PP2C family (**Figure S6**). The diversity of its output domains identified this 6TM region as a novel integral membrane sensor domain, which was named MASE10. The presence of conserved Cys, Ser, and Trp residues in the periplasmic loops of MASE10, as well as conservation of aromatic residues and four charged (Arg, Asp, Lys, and another Asp) residues in the predicted TM segments, suggest highly specific interaction of MASE10 with its ligand(s). In all three classes of signaling proteins, MASE10 domain was followed by the signal-transducing HAMP domain (80)(Figure S6).

MASE11. Sequence analysis of *Shewanella oneidensis* protein SO1387 revealed another conserved integral membrane domain (**Figure S7**). It contained 7 predicted TM segments but, as in MASE9, two predicted TM segments were too short to cross the membrane and formed a likely intramembrane hairpin. This domain was named MASE11. In addition to its combination with the HD-GYP domain, MASE11 was found in association with PAS, GAF, and GGDEF–EAL domains, as well as in sensor histidine kinases with various domain architectures (Figure S7). So far, no MASE10-containing protein has been experimentally characterized.

MASE12. A recent study of the roles of c-di-GMP in *Bacillus anthracis* demonstrated the involvement of the HD-GYP-containing protein Bas0914 (GenBank accession number [AAT53240.1](https://www.ncbi.nlm.nih.gov/nuccore/AAT53240.1)) in the regulation of bacterial swarming, albeit not in sporulation, germination, or toxin production (81). This protein, referred to as CdgM, combines HD-GYP with a conserved N-terminal integral membrane domain, hereafter MASE12 domain (**Figure S8**). The MASE12-HD-GYP domain combination is encoded in a number of bacilli but has not been seen outside of *Bacillales*. MASE12 was also found in bacillar histidine kinases, including BA5029, previously identified as one of the sporulation kinases of *B. anthracis* (82). Despite the relatively narrow phylogenetic distribution of MASE12, its presence in (predicted) c-di-GMP phosphodiesterases and histidine kinases indicates that it is a novel sensory

domain. The presence of conserved aromatic residues in each of the transmembrane segments of MASE12 suggests that its ligand could be a molecule with one or more aromatic rings.

Regulatory roles of HD-GYP-containing proteins

HD-GYP domains participate in bacterial regulatory cascades in two principal ways: by hydrolyzing c-di-GMP (and/or cGAMP) and by participating in complex protein-protein interactions (24, 25). Like GGDEF and EAL, the other two domains involved in c-di-GMP metabolism, HD-GYPs are rarely seen in a stand-alone form. As shown in Table 3 and can be seen in domain architecture charts of HD and HD_5 domains in Pfam, HD-GYPs are typically found in at the C-termini of multidomain proteins, fused to various N-terminal domains. By analogy with other signal transduction proteins that combine N-terminal sensor domains with the C-terminal output domain(s), such as histidine kinase, adenylate cyclase, GGDEF, EAL, or GGDEF-EAL pair, N-terminal (sensor) domains are expected to modulate the c-di-GMP phosphodiesterase activity of the downstream HD-GYP domains.

In those HD-GYP domains that are linked to extracytoplasmic (dCache, PBPb, CHASE2) ligand-binding sensor domains (Table 3), c-di-GMP hydrolase activity could be directly controlled by the outside signals. To the best of our knowledge, no such protein has been experimentally characterized so far. Fusions of HD-GYP with PAS and GAF domains likely control c-di-GMP levels in response to the intracellular signals, including the levels of AMP, GMP, the redox status of the cell, and other parameters. One such protein, PmGH, has been structurally characterized (45), but the ligands of its N-terminal GAF domain have not been identified. In another protein of the GAF–HD-GYP domain architecture, Tlr0485 from *Thermosynechococcus vulcanus* (GenBank accession number [BAC08037.1](#)), the c-di-GMP phosphodiesterase activity of the HD-GYP domain was dependent upon cAMP binding to its GAF domain (83). Another example of such regulation is the protein FV185_09380 (GenBank accession number [KXW56972.1](#)) from the betaproteobacterium *Ferroplasma* sp., whose N-terminal hemerythrin domain (PF01814 in Pfam) was shown to function as a redox and oxygen sensor that controls the activity of the downstream HD-GYP domain (84). While there is little doubt that the same regulatory mechanisms work in other HD-GYP-containing proteins, they remain to be directly demonstrated.

The widespread RpfG (REC–HD-GYP) family of response regulators (Table 3) includes both catalytically active and inactive forms. Several RpfG family proteins have been shown to hydrolyze c-di-GMP (34, 36, 38, 39) and their phosphodiesterase activity was dependent on the phosphorylated state of their N-terminal REC domains. Thus, along with the response regulators of the WspR (REC–GGDEF), PleD (REC–REC–GGDEF), and PvrR (REC–EAL) families that catalyze respectively, synthesis and hydrolysis of c-di-GMP, RpfG family proteins require their cognate sensor histidine kinases (or acetyl phosphate) for activity and thereby put cellular levels of c-di-GMP under the control of the two-component signal transduction systems. This is yet another way for extracellular signals to control multiple c-di-GMP-dependent functions.

Catalytically inactive proteins of the RpfG family may still control synthesis and secretion of extracellular enzymes and exopolysaccharides through protein-protein interactions. One example is the eponymous RpfG from *X. campestris* (65, 66, 85, 86), which does not appear to have significant c-

di-GMP phosphodiesterase activity. Another protein of the RpfG family, *Shewanella oneidensis* HnoD (SO_2541, GenBank: [AAN55571.1](#)) with degenerate active motifs (HD→SE, HHE→QLE, see Figure S1), did not hydrolyze c-di-GMP by itself and inhibited c-di-GMP hydrolysis by the PvrR family response regulator HnoB (REC–EAL domain combination (34)). Another example is protein ECA3548 (GenBank: [CAG76446](#)) from *Pectobacterium atrosepticum*, in which an inactivated HD-GYP domain with altered active-site motifs (HD→HT, HHE→ANE, see Figure S1) is fused to a YcgR-like combination of N-terminal PilZ (aka YcgR, [PF07317](#)) and canonical PilZ ([PF07238](#)) domains (87). Such a domain combination should be able to bind c-di-GMP but not hydrolyze it, which is consistent with the absence of either c-di-GMP or pGpG hydrolysis by this protein (40). Overexpression of ECA3548 led to increased c-di-GMP levels and affected cell motility (40). There is little doubt that the list of such catalytically inactive HD-GYP domain proteins that are involved in regulatory interactions will continue to grow

CONCLUDING REMARKS

This study was prompted by the absence of clear data on the distribution of the HD-GYP domains in the public databases. Fortunately, analysis of the domain architectures of HD and HD-GYP domain proteins revealed a clear distinction between them: ‘classical’ HD-domains were typically associated with either house-keeping enzymatic domains, such as RNA polymerase, RNaseY, helicases, and various nucleotidyltransferases, or with defense systems (Cas3 of the CRISPR-Cas system) (27-30). By contrast, HD-GYP domains were always associated with other signaling domains, either sensory (ligand-binding) or signal transduction domains (Table 3). Moreover, certain widespread domain architectures appeared to be limited to the HD-GYP domains. Indeed, CD-search of various proteins combining the HD-type domains with REC, GAF, PBPb, dCache, CHASE2, or DUF3391 domains showed that almost all of them actually contained HD-GYP domains. The only enzymes with the ‘classical’ HD domains with a role in signal transduction were the (p)ppGpp-hydrolyzing enzymes of the RelA/Spot Homologue (RSH) family and c-di-AMP-hydrolyzing enzymes of the PgpH family (88, 89), see Figure 1. It should be noted, however, that many HD-GYP-containing proteins have undefined N-terminal domains; such proteins comprise the most common domain architectures for both HD ([PF01966](#)) and HD_5 ([PF13487](#)) entries in Pfam (48). This work delineated three such domains, MASE9, MAS10, and MASE11 (Figures S5, S6, and S7), but there are definitely many more.

Compared to the ‘classical’ HD domains, HD-GYP domains show conservation of several additional residues, including the eponymous GYP motif (Figure 1), whose central Tyr residue binds a nonbridging oxygen atom bond in the central ribose-phosphate ring of c-di-GMP, see Figure 2. All these residues play a role in metal and/or c-di-GMP binding (Table 1). The involvement of multiple residues in c-di-GMP binding is not unique for the HD-GYP domain and has been seen in such c-di-GMP binding proteins as *V. cholerae* MshEN (90) or *Streptomyces venezuelae* RsiG (91, 92). It is important to note partial conservation of the metal-binding Glu residues in the N-terminal loop (E185 in PmGH, Figure 1). Phylogenetic studies of the HD-GYP domains reveal a clear split between those with binuclear (lacking this Glu residue) and trinuclear (with this Glu) metal sites, with the latter group apparently capable of both steps of c-di-GMP hydrolysis, first to pGpG and then to GMP (36, 45, 46, 58). Remarkably, many organisms, including *V. cholerae*, *P. aeruginosa*, and *D. vulgaris*, appear to encode members of both families (38, 39), see Figure S1. In the emerging pathogen *Shewanella algae*, four out of seven HD-GYP-containing proteins belong to one family and three to another (Martín-

Rodríguez. *et al.* submitted). Thus, if pGpG has its own functional role(s), as suggested by some authors (24, 37), some HD-GYP-containing proteins could play a role in hydrolyzing it as well.

Since the HD-GYP logos, presented in Figure 1B and Figure S2, are obtained from sequences that were retrieved from BLAST searches with catalytically active enzymes as queries, they may introduce a certain bias, showing the key amino acid residues as being more conserved than they are. Indeed, our study revealed a substantial number of HD-GYP domains with various substitutions in the active site residues. Such HD-GYP domains are expected to lack catalytic activity but may still retain the ability to bind c-di-GMP and/or participate in various protein-protein interactions. While proteins with GGDEF and EAL domains often participate in local signaling (13-15), there are very few reports of HD-GYP domain proteins acting this way (34, 40, 66). The full extent of such regulation remains to be determined, it is expected to be most pronounced in organisms where HD-GYP is the dominant c-di-GMP phosphodiesterase (Table 2). We hope that the ability to clearly define the HD-GYP domains, whether fully conserved or not, based on the alignment in Figure S1, would allow further insight into the evolution and functional divergence of the HD-GYP-containing proteins.

In summary, HD-GYP domains represent a specialized, highly conserved, widespread, and apparently ancient branch of the vast HD superfamily. Sequence conservation and the narrow functional repertoire of HD-GYP domains can be contrasted with the extreme divergence of the 'classical' HD domains that led to the emergence of numerous enzymes with distinct catalytic activities (26, 29, 30), all but two of them unrelated to signaling (88, 89). The HD-GYP and EAL domain counts, presented in Tables 2 and S2, show a wide distribution of the HD-GYP domains and their presence as the prevalent c-di-GMP hydrolases in many diverse bacteria. We hope that an updated description of the HD-GYP domain, its properties and domain architectures, presented in this work, would stimulate further studies of its regulatory roles, including the spectrum of ligands that affect its c-di-GMP phosphodiesterase activity and/or protein-binding properties.

MATERIALS AND METHODS

Identification of HD-GYP domain proteins. Putative HD-GYP domains were identified by searching protein sequence databases with PSI-BLAST (56) and HMMER (57) using the *Aquifex aeolicus* protein Aq_2027 (GenBank accession number [AAC07788.1](https://www.ncbi.nlm.nih.gov/nuclot/AAC07788.1)) and the *Thermotoga maritima* protein TM0186 (GenBank accession number [AAD35279.1](https://www.ncbi.nlm.nih.gov/nuclot/AAD35279.1)) (3) as queries. Domain architectures of the identified proteins were analyzed using CD-search (50, 93) and/or HHpred (73). These domain architectures were further compared with those in Pfam (48), SMART (49), and InterPro (51) databases. To evaluate the presence of HD-GYP domains in various Pfam domain architectures with HD ([PF01966](https://www.pfam.org/entry/1966)) and HD_5 ([PF13487](https://www.pfam.org/entry/13487)) domain entries, the lists of proteins with respective architectures were downloaded from the Pfam website. Their GenBank/EMBL/DDBJ accession numbers were either taken from the respective protein defines or retrieved using the UniProt's Retrieve/ID mapping (<https://www.uniprot.org/uploadlists/>) tool and used in the CD-search (93) in the batch mode (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>). The resulting hits against [cd00077](https://www.ncbi.nlm.nih.gov/nuclot/cd00077), [COG2206](https://www.ncbi.nlm.nih.gov/nuclot/COG2206), [COG3437](https://www.ncbi.nlm.nih.gov/nuclot/COG3437), [PF01966](https://www.ncbi.nlm.nih.gov/nuclot/PF01966), [PF13487](https://www.ncbi.nlm.nih.gov/nuclot/PF13487), and [SM00471](https://www.ncbi.nlm.nih.gov/nuclot/SM00471) domain models were individually checked for the alignment length and the presence of the HHEXh[D/N]GxGYP sequence motif (*h*, hydrophobic

residue; x, any residue). The list of experimentally characterized HD-GYP-containing proteins was compiled from literature sources; the presence of HD-GYP domain was verified with CD-search.

Structure-based alignment of HD-GYP domains. Four distinct structures of the HD-GYP domains were retrieved from the Protein DataBank (PDB) and aligned using DALI (94) and VAST (95) servers. These alignments were then compared to that presented in the original description of the HD-GYP domain (3) and further aligned against [cd00077](#), [COG2206](#), [COG3437](#), [PF01966](#), [PF13487](#), and [SM00471](#) domain models (Figure 1A). Experimentally characterized HD-GYP domains and those from phylogenetically diverse bacteria were added to the expanded alignment (Figure S1) based on their CD-search hits.

Analysis of HD-GYP-containing domain architectures. The lists of proteins with specific domain architectures were taken from Pfam's "Domain organisation" links for HD and HD_5 domain entries (48) and/or retrieved from CD-search outputs using CDART (96). For the HD-GYP sequence logo, 325 best hits from the PSI-BLAST (56) search of Aq_2027 sequence against the Pfam's set of proteins with GAF-HD-GYP domain architecture were retrieved, aligned using the "Flat query-anchored with letters for identities" option of BLAST output, trimmed to fit the sequence consensus shown on Figure 1, and submitted to the WebLogo (97) tool. The same procedure has been used to generate sequence logos of the RpfG (REC-HD-GYP) and Bd1817 (DUF3391-HD-GYP) families.

HD-GYP and EAL domain counts. The numbers of HD-GYP and EAL domains encoded in specific bacterial genomes were taken from the COG database (52), using combined numbers from COG2206 and COG3437 for HD-GYP and combined numbers from COG2200, COG4943, and COG5001 for the EAL domain. The numbers were verified using the CD-search tool in the batch mode and checking the resulting hits for the alignment length and the presence of the HHEXh[D/N]GXGYP sequence motif.

Analysis of DUF and MASE domains. The alignments of two DUF and three MASE domains were constructed from the PSI-BLAST (56) and HMMER (57) outputs and, where appropriate, manually edited to fit the sequence consensus. The secondary structure of the DUF3391 domain was taken from the PDB entry 3TM8 (44); secondary structure of DUF3369 was predicted using HHPred (73) and JPred (98). For MASE domains, transmembrane segments and their membrane topology were predicted with TMHMM2.0 (99) and compared with UniProt (100) and InterPro (51) entries for the selected proteins.

ACKNOWLEDGEMENTS

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. We thank Prof. Dr. Natalia Tschowri (Hannover) for helpful comments.

REFERENCES

1. Ross P, Weinhouse H, Aloni Y, Michaeli D, Weinberger-Ohana P, Mayer R, Braun S, de Vroom E, van der Marel GA, van Boom JH, Benziman M. 1987. Regulation of cellulose synthesis in *Acetobacter xylinum* by cyclic diguanylic acid. *Nature* 325:279-281; doi: 10.1038/325279a0.
2. Tal R, Wong HC, Calhoon R, Gelfand DH, Fear AL, Volman G, Mayer R, Ross P, Amikam D,

- Weinhouse H, Cohen A, Sapir S, Ohana P, Benziman M. 1998. Three *cdg* operons control cellular turnover of cyclic di-GMP in *Acetobacter xylinum*: genetic organization and occurrence of conserved domains in isoenzymes. J Bacteriol 180:4416-4425; doi: 10.1128/JB.180.17.4416-4425.1998.
3. Galperin MY, Natale DA, Aravind L, Koonin EV. 1999. A specialized version of the HD hydrolase domain implicated in signal transduction. J Mol Microbiol Biotechnol 1:303-305.
4. Galperin MY, Nikolskaya AN, Koonin EV. 2001. Novel domains of the prokaryotic two-component signal transduction systems. FEMS Microbiol Lett 203:11-21; doi: 10.1111/j.1574-6968.2001.tb10814.x.
5. Galperin MY. 2004. Bacterial signal transduction network in a genomic perspective. Environ Microbiol 6:552-567; doi: 10.1111/j.1462-2920.2004.00633.x.
6. Römling U, Gomelsky M, Galperin MY. 2005. C-di-GMP: the dawning of a novel bacterial signalling system. Mol Microbiol 57:629-639; doi: 10.1111/j.1365-2958.2005.04697.x.
7. Römling U, Galperin MY, Gomelsky M. 2013. Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. Microbiol Mol Biol Rev 77:1-52; doi: 10.1128/MMBR.00043-12.
8. Conner JG, Zamorano-Sanchez D, Park JH, Sondermann H, Yildiz FH. 2017. The ins and outs of cyclic di-GMP signaling in *Vibrio cholerae*. Curr Opin Microbiol 36:20-29; doi: 10.1016/j.mib.2017.01.002.
9. Jenal U, Reinders A, Lori C. 2017. Cyclic di-GMP: second messenger extraordinaire. Nat Rev Microbiol 15:271-284; doi: 10.1038/nrmicro.2016.190.
10. Su J, Zou X, Huang L, Bai T, Liu S, Yuan M, Chou SH, He YW, Wang H, He J. 2016. DgcA, a diguanylate cyclase from *Xanthomonas oryzae* pv. *oryzae* regulates bacterial pathogenicity on rice. Sci Rep 6:25978; doi: 10.1038/srep25978.
11. Hall CL, Lee VT. 2018. Cyclic-di-GMP regulation of virulence in bacterial pathogens. Wiley Interdiscip Rev RNA 9:e1454; doi: 10.1002/wrna.1454.
12. Valentini M, Filloux A. 2019. Multiple roles of c-di-GMP signaling in bacterial pathogenesis. Annu Rev Microbiol 73:387-406; doi: 10.1146/annurev-micro-020518-115555.
13. Sarenko O, Klauck G, Wilke FM, Pfiffer V, Richter AM, Herbst S, Kaever V, Hengge R. 2017. More than enzymes that make or break cyclic di-GMP - local signaling in the interactome of GGDEF/EAL domain proteins of *Escherichia coli*. mBio 8:e01639-01617; doi: 10.1128/mBio.01639-17.
14. Richter AM, Possling A, Malysheva N, Yousef KP, Herbst S, von Kleist M, Hengge R. 2020. Local c-di-GMP signaling in the control of synthesis of the *E. coli* biofilm exopolysaccharide pEtN-cellulose. J Mol Biol 432:4576-4595; doi: 10.1016/j.jmb.2020.06.006.
15. Hengge R. 2021. High-specificity local and global c-di-GMP signaling. Trends Microbiol 29:993-1003; doi: 10.1016/j.tim.2021.02.003.
16. Chan C, Paul R, Samoray D, Amiot NC, Giese B, Jenal U, Schirmer T. 2004. Structural basis of activity and allosteric control of diguanylate cyclase. Proc Natl Acad Sci USA 101:17084-17089; doi: 10.1073/pnas.0406134101.
17. Schirmer T, Jenal U. 2009. Structural and mechanistic determinants of c-di-GMP signalling. Nat Rev Microbiol 7:724-735; doi: 10.1038/nrmicro2203.
18. Schirmer T. 2016. C-di-GMP synthesis: Structural aspects of evolution, catalysis and regulation. J Mol Biol 428:3683-3701; doi: 10.1016/j.jmb.2016.07.023.
19. Minasov G, Padavattan S, Shuvalova L, Brunzelle JS, Miller DJ, Basle A, Massa C, Collart FR, Schirmer T, Anderson WF. 2009. Crystal structures of Ykul and its complex with second messenger cyclic di-GMP suggest catalytic mechanism of phosphodiester bond cleavage by EAL domains. J Biol Chem 284:13174-13184; doi: 10.1074/jbc.M808221200.
20. Tchigvintsev A, Xu X, Singer A, Chang C, Brown G, Proudfoot M, Cui H, Flick R, Anderson WF, Joachimiak A, Galperin MY, Savchenko A, Yakunin AF. 2010. Structural insight into the mechanism of c-di-GMP hydrolysis by EAL domain phosphodiesterases. J Mol Biol 402:524-538; doi: 10.1016/j.jmb.2010.07.050.

21. Krasteva PV, Giglio KM, Sondermann H. 2012. Sensing the messenger: The diverse ways that bacteria signal through c-di-GMP. *Protein Sci* 21:929-948; doi: 10.1002/pro.2093.
22. Wigren E, Liang ZX, Römling U. 2014. Finally! The structural secrets of a HD-GYP phosphodiesterase revealed. *Mol Microbiol* 91:1-5; doi: 10.1111/mmi.12463.
23. Römling U, Liang ZX, Dow JM. 2017. Progress in understanding the molecular basis underlying functional diversification of cyclic dinucleotide turnover proteins. *J Bacteriol* 199:e00790-16; doi: 10.1128/JB.00790-16.
24. Rinaldo S, Paiardini A, Paone A, Cutruzzola F, Giardina G. 2020. Structure and function of HD-GYP phosphodiesterases, p 65-78. *In* Guilian N, Chou S-H, Römling U, Lee VT (ed), *Microbial Cyclic Di-Nucleotide Signaling*, 2015/02/19 ed. Springer, Cham.
25. Cutruzzola F, Paiardini A, Rossi CS, Spizzichino S, Paone A, Giardina G, Rinaldo S. 2022. A conserved scaffold with heterogeneous metal ion binding site: the multifaceted example of HD-GYP proteins. *Coord Chem Rev* 450:214228; doi: 10.1016/j.ccr.2021.214228.
26. Aravind L, Koonin EV. 1998. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem Sci* 23:469-472; doi: 10.1016/s0968-0004(98)01293-6.
27. Beloglazova N, Petit P, Flick R, Brown G, Savchenko A, Yakunin AF. 2011. Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J* 30:4616-4627; doi: 10.1038/emboj.2011.377.
28. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. 2011. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 9:467-477; doi: 10.1038/nrmicro2577.
29. Galperin MY, Koonin EV. 2012. Divergence and convergence in enzyme evolution. *J Biol Chem* 287:21-28; doi: 10.1074/jbc.R111.241976.
30. Langton M, Sun S, Ueda C, Markey M, Chen J, Paddy I, Jiang P, Chin N, Milne A, Pandelia ME. 2020. The HD-domain metalloprotein superfamily: An apparent common protein scaffold with diverse chemistries. *Catalysts* 10:1191; doi: 10.3390/catal10101191.
31. Ryan RP, Fouhy Y, Lucey JF, Crossman LC, Spiro S, He YW, Zhang LH, Heeb S, Camara M, Williams P, Dow JM. 2006. RETRACTED: Cell-cell signaling in *Xanthomonas campestris* involves an HD-GYP domain protein that functions in cyclic di-GMP turnover. *Proc Natl Acad Sci USA* 103:6712-6717; doi: 10.1073/pnas.0600345103. (Retraction published 15 Aug 2017, *Proc Natl Acad Sci USA* 114:E7031).
32. Hammer BK, Bassler BL. 2009. Distinct sensory pathways in *Vibrio cholerae* El Tor and classical biotypes modulate c-di-GMP levels to control biofilm formation. *J Bacteriol* 191:169-177; doi: 10.1128/JB.01307-08.
33. Sultan SZ, Pitzer JE, Boquoy T, Hobbs G, Miller MR, Motaleb MA. 2011. Analysis of the HD-GYP domain cyclic dimeric GMP phosphodiesterase reveals a role in motility and the enzootic life cycle of *Borrelia burgdorferi*. *Infect Immun* 79:3273-3283; doi: 10.1128/IAI.05153-11.
34. Plate L, Marletta MA. 2012. Nitric oxide modulates bacterial biofilm formation through a multicomponent cyclic-di-GMP signaling network. *Mol Cell* 46:449-460; doi: 10.1016/j.molcel.2012.03.023.
35. Miner KD, Klose KE, Kurtz DM, Jr. 2013. An HD-GYP cyclic di-guanosine monophosphate phosphodiesterase with a non-heme diiron-carboxylate active site. *Biochemistry* 52:5329-5331; doi: 10.1021/bi4009215.
36. Miner KD, Kurtz DM, Jr. 2016. Active site metal occupancy and cyclic di-GMP phosphodiesterase activity of *Thermotoga maritima* HD-GYP. *Biochemistry* 55:970-979; doi: 10.1021/acs.biochem.5b01227.
37. Stelitano V, Giardina G, Paiardini A, Castiglione N, Cutruzzola F, Rinaldo S. 2013. C-di-GMP hydrolysis by *Pseudomonas aeruginosa* HD-GYP phosphodiesterases: analysis of the reaction mechanism and novel roles for pGpG. *PLoS One* 8:e74920; doi: 10.1371/journal.pone.0074920.
38. McKee RW, Kariisa A, Mudrak B, Whitaker C, Tamayo R. 2014. A systematic analysis of the in vitro and in vivo functions of the HD-GYP domain proteins of *Vibrio cholerae*. *BMC Microbiol*

- 14:272; doi: 10.1186/s12866-014-0272-9.
39. Rajeev L, Luning EG, Altenburg S, Zane GM, Baidoo EE, Catena M, Keasling JD, Wall JD, Fields MW, Mukhopadhyay A. 2014. Identification of a cyclic-di-GMP-modulating response regulator that impacts biofilm formation in a model sulfate reducing bacterium. *Front Microbiol* 5:382; doi: 10.3389/fmicb.2014.00382.
40. Tan H, West JA, Ramsay JP, Monson RE, Griffin JL, Toth IK, Salmond GPC. 2014. Comprehensive overexpression analysis of cyclic-di-GMP signalling proteins in the phytopathogen *Pectobacterium atrosepticum* reveals diverse effects on motility and virulence phenotypes. *Microbiology (Reading)* 160:1427-1439; doi: 10.1099/mic.0.076828-0.
41. Gao J, Tao J, Liang W, Zhao M, Du X, Cui S, Duan H, Kan B, Su X, Jiang Z. 2015. Identification and characterization of phosphodiesterases that specifically degrade 3'3'-cyclic GMP-AMP. *Cell Res* 25:539-550; doi: 10.1038/cr.2015.40.
42. Deng MJ, Tao J, E C, Ye ZY, Jiang Z, Yu J, Su XD. 2018. Novel mechanism for cyclic dinucleotide degradation revealed by structural studies of *Vibrio* phosphodiesterase V-cGAP3. *J Mol Biol* 430:5080-5093; doi: 10.1016/j.jmb.2018.10.010.
43. Wright TA, Jiang L, Park JJ, Anderson WA, Chen G, Hallberg ZF, Nan B, Hammond MC. 2020. Second messengers and divergent HD-GYP phosphodiesterases regulate 3',3'-cGAMP signaling. *Mol Microbiol* 113:222-236; doi: 10.1111/mmi.14412.
44. Lovering AL, Capeness MJ, Lambert C, Hobley L, Sockett RE. 2011. The structure of an unconventional HD-GYP protein from *Bdellovibrio* reveals the roles of conserved residues in this class of cyclic-di-GMP phosphodiesterases. *mBio* 2: e00163-11; doi: 10.1128/mBio.00163-11.
45. Bellini D, Caly DL, McCarthy Y, Bumann M, An SQ, Dow JM, Ryan RP, Walsh MA. 2014. Crystal structure of an HD-GYP domain cyclic-di-GMP phosphodiesterase reveals an enzyme with a novel trinuclear catalytic iron centre. *Mol Microbiol* 91:26-38; doi: 10.1111/mmi.12447.
46. Rinaldo S, Paiardini A, Stelitano V, Brunotti P, Cervoni L, Fernicola S, Protano C, Vitali M, Cutruzzola F, Giardina G. 2015. Structural basis of functional diversification of the HD-GYP domain revealed by the *Pseudomonas aeruginosa* PA4781 protein, which displays an unselective bimetallic binding site. *J Bacteriol* 197:1525-1535; doi: 10.1128/JB.02606-14.
47. Anantharaman V, Balaji S, Aravind L. 2006. The signaling helix: a common functional theme in diverse signaling proteins. *Biol Direct* 1:25; doi: 10.1186/1745-6150-1-25.
48. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res* 49:D412-D419; doi: 10.1093/nar/gkaa913.
49. Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res* 49:D458-D460; doi: 10.1093/nar/gkaa937.
50. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* 48:D265-D268; doi: 10.1093/nar/gkz991.
51. Blum M, Chang HY, Chuguransky S, Grego T, Kandasamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 49:D344-D354; doi: 10.1093/nar/gkaa977.
52. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 49:D274-D281; doi: 10.1093/nar/gkaa1018.
53. Sigrist CJ, de Castro E, Cerutti L, Cuče BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2013. New and continuing developments at PROSITE. *Nucleic Acids Res* 41:D344-347; doi:

- 10.1093/nar/gks1067.
54. Galperin MY. 2006. Structural classification of bacterial response regulators: diversity of output domains and domain combinations. *J Bacteriol* 188:4169-4182; doi: 10.1128/JB.01887-05.
55. Lee HM, Liao CT, Chiang YC, Chang YY, Yeh YT, Du SC, Hsiao YM. 2016. Characterization of genes encoding proteins containing HD-related output domain in *Xanthomonas campestris* pv. *campestris*. *Antonie Van Leeuwenhoek* 109:509-522; doi: 10.1007/s10482-016-0656-y.
56. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zheng Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST - A new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402; doi: 10.1093/nar/25.17.3389.
57. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. *Nucleic Acids Res* 43:W30-W38; doi: 10.1093/nar/gkv397.
58. Sun S, Pandelia ME. 2020. HD-[HD-GYP] phosphodiesterases: Activities and evolutionary diversification within the HD-GYP family. *Biochemistry* 59:2340-2350; doi: 10.1021/acs.biochem.0c00257.
59. Chou SH, Galperin MY. 2016. Diversity of c-di-GMP-binding proteins and mechanisms. *J Bacteriol* 198:32-46; doi: 10.1128/JB.00333-15.
60. Al-Bassam MM, Haist J, Neumann SA, Lindenberg S, Tschowri N. 2018. Expression patterns, genomic conservation and input into developmental regulation of the GGDEF/EAL/HD-GYP domain proteins in *Streptomyces*. *Front Microbiol* 9:2524; doi: 10.3389/fmicb.2018.02524.
61. Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S, Schneeberger K. 2016. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci USA* 113:E4052-4060; doi: 10.1073/pnas.1607532113.
62. Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, Schmidt TSB, Bork P. 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* 22:178; doi: 10.1186/s13059-021-02393-0.
63. Zhou Z, Liu Y, Xu W, Pan J, Luo ZH, Li M. 2020. Genome- and community-level interaction insights into carbon utilization and element cycling functions of *Hydrothermarchaeota* in hydrothermal sediment. *mSystems* 5:e00795-19; doi: 10.1128/mSystems.00795-19.
64. Lebedinsky AV, Mardanov AV, Kublanov IV, Gumerov VM, Beletsky AV, Perevalova AA, Bidzhieva S, Bonch-Osmolovskaya EA, Skryabin KG, Ravin NV. 2014. Analysis of the complete genome of *Fervidococcus fontis* confirms the distinct phylogenetic position of the order Fervidococcales and suggests its environmental function. *Extremophiles* 18:295-309; doi: 10.1007/s00792-013-0616-7.
65. Slater H, Alvarez-Morales A, Barber CE, Daniels MJ, Dow JM. 2000. A two-component system involving an HD-GYP domain protein links cell-cell signalling to pathogenicity gene expression in *Xanthomonas campestris*. *Mol Microbiol* 38:986-1003; doi: 10.1046/j.1365-2958.2000.02196.x.
66. Andrade MO, Alegria MC, Guzzo CR, Docena C, Rosa MC, Ramos CH, Farah CS. 2006. The HD-GYP domain of RpfG mediates a direct linkage between the Rpf quorum-sensing pathway and a subset of diguanylate cyclase proteins in the phytopathogen *Xanthomonas axonopodis* pv *citri*. *Mol Microbiol* 62:537-551.
67. Zhang Y, Wei C, Jiang W, Wang L, Li C, Wang Y, Dow JM, Sun W. 2013. The HD-GYP domain protein RpfG of *Xanthomonas oryzae* pv. *oryzicola* regulates synthesis of extracellular polysaccharides that contribute to biofilm formation and virulence on rice. *PLoS One* 8:e59428; doi: 10.1371/journal.pone.0059428.
68. Yaryura PM, Conforte VP, Malamud F, Roeschlin R, de Pino V, Castagnaro AP, McCarthy Y, Dow JM, Marano MR, Vojnov AA. 2015. XbmR, a new transcription factor involved in the regulation of chemotaxis, biofilm formation and virulence in *Xanthomonas citri* subsp. *citri*. *Environ Microbiol* 17:4164-4176; doi: 10.1111/1462-2920.12684.
69. Wei C, Jiang W, Zhao M, Ling J, Zeng X, Deng J, Jin D, Dow JM, Sun W. 2016. A systematic analysis

- of the role of GGDEF-EAL domain proteins in virulence and motility in *Xanthomonas oryzae* pv. *oryzicola*. Sci Rep 6:23769; doi: 10.1038/srep23769.
70. Li K, Xu G, Wang B, Wu G, Hou R, Liu F. 2021. The predatory soil bacterium *Lysobacter* reprograms quorum sensing system to regulate antifungal antibiotic production in a cyclic-di-GMP-independent manner. Commun Biol 4:1131; doi: 10.1038/s42003-021-02660-7.
71. Galperin MY. 2010. Diversity of structure and function of response regulator output domains. Curr Opin Microbiol 13:150-159; doi: 10.1016/j.mib.2010.01.005.
72. Galperin MY. 2018. What bacteria want. Environ Microbiol 20:4221-4229; doi: 10.1111/1462-2920.14398.
73. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. J Mol Biol 430:2237-2243; doi: 10.1016/j.jmb.2017.12.007.
74. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci USA 117:1496-1503; doi: 10.1073/pnas.1914677117.
75. Nikolskaya AN, Mulikidjanian AY, Beech IB, Galperin MY. 2003. MASE1 and MASE2: two novel integral membrane sensory domains. J Mol Microbiol Biotechnol 5:11-16; doi: 10.1159/000068720.
76. Galperin MY, Nikolskaya AN. 2007. Identification of sensory and signal-transducing domains in two-component signaling systems. Methods Enzymol 422:47-74; doi: 10.1016/S0076-6879(06)22003-2.
77. Hengge R, Galperin MY, Ghigo JM, Gomelsky M, Green J, Hughes KT, Jenal U, Landini P. 2016. Systematic nomenclature for GGDEF and EAL domain-containing cyclic di-GMP turnover proteins of *Escherichia coli*. J Bacteriol 198:7-11; doi: 10.1128/JB.00424-15.
78. Koestler BJ, Waters CM. 2014. Bile acids and bicarbonate inversely regulate intracellular cyclic di-GMP in *Vibrio cholerae*. Infect Immun 82:3002-3014; doi: 10.1128/IAI.01664-14.
79. Roelofs KG, Jones CJ, Helman SR, Shang X, Orr MW, Goodson JR, Galperin MY, Yildiz FH, Lee VT. 2015. Systematic identification of cyclic-di-GMP binding proteins in *Vibrio cholerae* reveals a novel class of cyclic-di-GMP-binding ATPases associated with type II secretion systems. PLoS Pathog 11:e1005232; doi: 10.1371/journal.ppat.1005232.
80. Aravind L, Ponting CP. 1999. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. FEMS Microbiol Lett 176:111-116; doi: 10.1111/j.1574-6968.1999.tb13650.x.
81. Hermanas TM, Subramanian S, Dann CE, 3rd, Stewart GC. 2021. Spore-associated proteins involved in c-di-GMP synthesis and degradation of *Bacillus anthracis*. J Bacteriol 203:e0013521; doi: 10.1128/JB.00135-21.
82. Brunsing RL, La Clair C, Tang S, Chiang C, Hancock LE, Perego M, Hoch JA. 2005. Characterization of sporulation histidine kinases of *Bacillus anthracis*. J Bacteriol 187:6972-6981; doi: 10.1128/JB.187.20.6972-6981.2005.
83. Enomoto G, Kamiya A, Okuda Y, Narikawa R, Ikeuchi M. 2020. Tlr0485 is a cAMP-activated c-di-GMP phosphodiesterase in a cyanobacterium *Thermosynechococcus*. J Gen Appl Microbiol 66:147-152; doi: 10.2323/jgam.2020.01.007.
84. Kitanishi K, Igarashi J, Matsuoka A, Unno M. 2020. Identification and characterization of a redox sensor phosphodiesterase from *Ferroplasma* sp. PN-J185 containing bacterial hemerythrin and HD-GYP domains. Biochemistry 59:983-991; doi: 10.1021/acs.biochem.0c00021.
85. Dow JM, Crossman L, Findlay K, He YQ, Feng JX, Tang JL. 2003. Biofilm dispersal in *Xanthomonas campestris* is controlled by cell-cell signaling and is required for full virulence to plants. Proc Natl Acad Sci USA 100:10995-11000; doi: 10.1073/pnas.1833360100.
86. Dow JM, Fouhy Y, Lucey JF, Ryan RP. 2006. The HD-GYP domain, cyclic di-GMP signaling, and bacterial virulence to plants. Mol Plant Microbe Interact 19:1378-1384.
87. Galperin MY, Chou SH. 2020. Structural conservation and diversity of PilZ-related domains. J

- Bacteriol 202:e00664-19; doi: 10.1128/JB.00664-19.
88. Atkinson GC, Tenson T, Hauryliuk V. 2011. The RelA/SpoT homolog (RSH) superfamily: distribution and functional evolution of ppGpp synthetases and hydrolases across the tree of life. PLoS One 6:e23479; doi: 10.1371/journal.pone.0023479.
89. Huynh TN, Luo S, Pensinger D, Sauer JD, Tong L, Woodward JJ. 2015. An HD-domain phosphodiesterase mediates cooperative hydrolysis of c-di-AMP to affect bacterial growth and virulence. Proc Natl Acad Sci USA 112:E747-E756; doi: 10.1073/pnas.1416485112.
90. Wang YC, Chin KH, Tu ZL, He J, Jones CJ, Sanchez DZ, Yildiz FH, Galperin MY, Chou SH. 2016. Nucleotide binding by the widespread high-affinity cyclic di-GMP receptor MshEN domain. Nat Commun 7:12481; doi: 10.1038/ncomms12481.
91. Gallagher KA, Schumacher MA, Bush MJ, Bibb MJ, Chandra G, Holmes NA, Zeng W, Henderson M, Zhang H, Findlay KC, Brennan RG, Buttner MJ. 2020. c-di-GMP arms an anti- σ to control progression of multicellular differentiation in *Streptomyces*. Mol Cell 77:586-599 e586; doi: 10.1016/j.molcel.2019.11.006.
92. Schumacher MA, Gallagher KA, Holmes NA, Chandra G, Henderson M, Kysela DT, Brennan RG, Buttner MJ. 2021. Evolution of a σ -(c-di-GMP)-anti- σ switch. Proc Natl Acad Sci USA 118:e2105447118; doi: 10.1073/pnas.2105447118.
93. Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32:W327-W331.
94. Holm L. 2020. Using Dali for protein structure comparison. Methods Mol Biol 2112:29-42; doi: 10.1007/978-1-0716-0270-6_3.
95. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH. 2014. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res 42:D297-D303; doi: 10.1093/nar/gkt1208.
96. Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: protein homology by domain architecture. Genome Res 12:1619-1623; doi: 10.1101/gr.278202.
97. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-1190; doi: 10.1101/gr.849004.
98. Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43:W389-W394; doi: 10.1093/nar/gkv332.
99. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567-580; doi: 10.1006/jmbi.2000.4315.
100. The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480-D489; doi: 10.1093/nar/gkaa1100.
101. Skotnicka D, Smaldone GT, Petters T, Trampari E, Liang J, Kaever V, Malone JG, Singer M, Sogaard-Andersen L. 2016. A minimal threshold of c-di-GMP is essential for fruiting body formation and sporulation in *Myxococcus xanthus*. PLoS Genet 12:e1006080; doi: 10.1371/journal.pgen.1006080.
102. Hogg T, Mechold U, Malke H, Cashel M, Hilgenfeld R. 2004. Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response. Cell 117:57-68; doi: 10.1016/s0092-8674(04)00260-0.

850 **Table 1. Functions of the conserved residues in the HD-GYP domain**

Residue number ^a in		Motif, position	Role in metal or c-di-GMP binding ^b
PmGH	Aq_2027		
Glu185	Tyr18	N/A	Binds Me2 and Me3
His189	His22	N/A	Binds Me3
His221	His54	<u>H</u> DhGK	Binds Me3
Asp222	Asp55	H <u>D</u> hGK	Binds Me1 and Me2
Lys225	Lys58	HDhG <u>K</u>	Binds Me2
Lys235	Lys68	<u>K</u> PG	Main chain carbonyl oxygen binds N2 of c-di-GMP Gua2
Gly237	Ala70	KP <u>G</u>	Main chain carbonyl oxygen binds N1 of c-di-GMP Gua2
His250	His83	N/A	Binds Me1
His276	His107	<u>H</u> HE	Binds Me1
His277	Gln108	H <u>H</u> E	Binds Me1
Asp278	Asp109	HHE <u>E</u>	
Gly284	Gly115	<u>G</u> YP	Stabilizes the 18-aa loop between α4 and α5
Tyr285	Tyr116	G <u>Y</u> P	Binds non-bridging O of c-di-GMP ribose-phosphate ring
Pro286	Pro117	GYP <u>P</u>	Stabilizes the 18-aa loop between α4 and α5
Asp305	Asp136	N/A	Binds Me3
Ala309	Ala140	<u>A</u> Lx ₃ RxhK	Hydrophobic interaction with c-di-GMP guanine base Gua1
Leu310	Met141	AL <u>L</u> x ₃ RxhK	Hydrophobic interaction with c-di-GMP guanine base Gua1
Arg314	Arg145	ALx ₃ <u>R</u> xhK	Binds N7 and O6 of c-di-GMP Gua1
Lys317	Lys146	ALx ₃ Rxh <u>K</u>	Binds O6 of c-di-GMP Gua1

851 ^a – See Figure 1, where PmGH is represented as 4MCW.

852 ^b – See Figure 2. Me1 and Me2 are metal (Fe²⁺) atoms in the binuclear metal center, presence
853 of Me3 makes it a trinuclear center. Me1 is the one closest to the GYP motif, Me3 is closest
854 to the HD motif (23, 45).

Table 2. Selected organisms with high numbers of HD-GYP domains

Organism ^a	Taxonomy ^b	Domain count	
		HD-GYP	EAL
Firmicutes			
<i>Acetobacterium woodii</i>	<i>Clostridia</i>	26	19
<i>Carboxydocella thermautotrophica</i>	<i>Clostridia</i>	18	2
<i>Desulfofundulus kuznetsovii</i>	<i>Clostridia</i>	19	1
<i>Halanaerobium hydrogeniformans</i>	<i>Clostridia</i>	22	2
<i>Natranaerobius thermophilus</i>	<i>Clostridia</i>	15	0
<i>Syntrophomonas wolfei</i>	<i>Clostridia</i>	17	0
<i>Limnochorda pilosa</i>	<i>Limnochordia</i>	6	1
<i>Methylophila anaerophila</i>	<i>Negativicutes</i>	17	3
Proteobacteria			
<i>Magnetospirillum magneticum</i>	<i>Alphaproteobacteria</i>	18	27
<i>Dechloromonas aromatica</i>	<i>Betaproteobacteria</i>	21	31
<i>Desulfatibacillum alkenivorans</i>	<i>Deltaproteobacteria</i>	12	4
<i>Desulfomonile tiedjei</i>	<i>Deltaproteobacteria</i>	7	0
<i>Bacteriovorax stolpii</i>	<i>Oligoflexia</i>	14	0
<i>Halobacteriovorax marinus</i>	<i>Oligoflexia</i>	12	0
Other phyla			
<i>Chloroherpeton thalassium</i>	<i>Chlorobi</i>	6	0
<i>Dictyoglomus thermophilum</i>	<i>Dictyoglomi</i>	8	0
<i>Candidatus Melainabacteria bacterium MEL.A1</i>	<i>Candidatus Melainabacteria</i>	5	0
<i>Nitrospira defluvii</i>	<i>Nitrospirae</i>	9	1
<i>Planctopirus limnophila</i>	<i>Planctomycetes</i>	5	0
<i>Thermanaerovibrio acidaminovorans</i>	<i>Synergistetes</i>	12	1
<i>Marinitoga piezophila</i>	<i>Thermotogae</i>	19	0

^a – The organisms and domain counts are taken from the COG database (52), COG2206 and COG3437 for the HD-GYP domain and COG2200, COG4943, and COG5001 for the EAL domain. Full organism names with strain numbers are listed in Table S2 in the supplemental materials. Domain numbers were verified using the CD-Search tool (50, 93) in the batch mode (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>). See Table S2 for the domain counts in 1,187 bacterial species.

^b – Class name for members of Firmicutes and Proteobacteria; for others, name of the phylum.

865 **Table 3. Domain architectures of selected HD-GYP-containing proteins**

Family name ^a , domain architecture ^b	Examples (locus tag, GenBank accession number)	References
RpfG family, REC – HD-GYP	XC2335 (CAB89843) TM0186 (AAD35279) ^c VC1348 (AAF94506) ^c DVU1181 (AAS95659) DVUA0086 (AAS94372) SO2541 (AAN55571) Gmet_3476 (ABB33684) ^d	(31, 34, 36, 38, 39, 41, 43)
Bd1817 family, DUF3391 – HD-GYP	Bd1817 (CAE79676) BB0374 (AAC66766) PA4108 (AAG07495) ^c TP0912 (AAC65863) TM1699 (AAC65727) VC2497 (AAF95639) XAC0350 (AAM35242)	(33, 37, 41, 44, 68)
PmGH family, GAF – HD-GYP	PERMA_0986 (ACO03551) Tlr0485 (BAC08037) TDE2659 (AAS13176) VC2340 (AAF95483)	(38, 45, 83)
VcGAP1 family, HD – HD-GYP	VCA0681 (AAF96581) ^{cd} VCA0931 (AAF96828) ^d Moth_2495 (ABC20777) ^c SO3491 (AAN56484)	(32, 38, 41-43, 58)
RpfX family, REC – PAS – HD-GYP	DVU2933 (AAS97405) SYN_01234 (ABC78345)	(39)
RpfY family, REC – DUF3369 – HD-GYP	DVU0722 (AAS95202) VCA0210 (AAF96122) ^{cd}	(38, 39, 41)
PmxA family, dCache – HAMP – HD-GYP	MXAN_2061 (ABF85902) ^c TM1682 (AAD36749)	(43, 101)
HdgA family, MASE9 – HD-GYP	TM1467 (AAD36535) Vnz_24090 (APE23802)	(60)
MASE10 – HAMP – HD-GYP	VC1295 (AAF94454) ^c Saut_1462 (ADN09509)	(38, 41)
MASE11 – HD-GYP	SO1387 (AAN54452) EW093_05120 (QEN04106)	
MASE12 – HD-GYP	Bas09147 (AAT53240)	(81)
HD-GYP – GGDEF	Aq_2027 (AAC07788) BLW93_04085 (OMH40677)	
PAS – HD-GYP	Dgeo_0663 (ABF44965) Glov_3421 (ACD97126)	
PBPb – HD-GYP	Dtur_1688 (ACK42960) TM1170 (AAD36245)	
CHASE2 – HD-GYP	Apau_0390 (EFQ22824) Dpep_1277 (EFC91303)	
Unc – HD-GYP	Bd2325 (CAE80147) TP_0764 (AAC65727) Vnz_24095 (APE23803)	(43, 60)

- 866 ^a – The families are named after their best-characterized member, if available.
- 867 ^b – Domain names are from the following entries in the Pfam or SMART databases: REC, PF00072
868 or SM00448; DUF3391, PF11871, see text and Figure S3; GAF, PF01590; PAS, PF00989;
869 DUF3369, PF11849, see text and Figure S4; dCache, PF02743; HAMP, PF00672; MASE9,
870 MASE10, and MASE11, see Figures S5, S6, and S7; GGDEF, PF00990; PBPb, PF00497 or
871 SM00062; CHASE2, PF05226; Unc, uncharacterized domain. Domains with multiple sequence
872 models, such as PAS_1..PAS_9 or GAF_1..GAF_3 were combined.
- 873 ^c – These proteins were able to hydrolyze c-di-GMP to GMP.
- 874 ^d – These proteins were able to hydrolyze cGAMP (41, 43).
- 875

A

```

-----HHHHHHHα1HHHHHHH-----HHHHHα2HHHHHH-----*-----HHHHHHHα3HHHHHHH-----HHα4H
Aq_2027 14 LRDAYTRGHTERGVSFAQLIGKELGLTSQEIETLLKLGGYTHDVGKVAIFDVLLKPAKLTPETVEYIMKLHVELGYEMVKDL--EIPWESL 101
3TM8 142 NTDKTISSHGVTVSTLSIALAQKLGIT2PKKTQLLTGLGALDHDYGHHSPLNLNQPLDSMSPEDLALWKKHPIDELAQKVQDK-KHFDQTVI 231
4MCW 181 FKDPETQNHIIIRVGLYAEILAREAGLDEEDVELVKLAAPMHDIKGVGIPDRVLLKPGKLNDEWEIMKKHTIYGYEILKGGDSRLQIAA 270
4R8Z 22 LRDNPRSRHLPRIERYVRLAEHLAA9PEAVDLLSKSALDHDICKVAVPDRVLLNPGQLDADTALLQGHFTRAGRDALASAERRFLRFAR 125
5Z7C 289 AKSPYTAGHSERVAVYTDLIARQLAISDADRIWLRAALIHDIKGLGVSNAILDKPGKLNDEWRAVQAHAAYTEQILYKL--SPFKTLA 376
BB_0374 131 LSNYHIIHSVNTAILTVAGNEMGLNNYKTVELCSIALDHDYGVGFIPDSVLLKPGKLTDEFTVMKEHPRIQAQALIAIA6SSFLAIAK 218
XC_2335 191 YRDGGTSAFLERMHVAGLVAEQGLSEEEVRIIEMAAPLHDMGKIAIPDSVLLKPGKLTDEEMNVMKRHPRIGYELLSGSQNRFIQVGA 280
TM_0186 165 YRDMETHRHFERVGVLSGRITAEEMGMSEVFTEIQFAAPDHDICKIGIPDRILLKPGILTPEFEIMKQHTITGFRILSRNSPILQLGA 254
PA_4108 150 LKTSDEYTYMHSVAVCALMTALAROL3DPLVREAGLAGLHDIKGMVDPDILNKPGLKTPDFEGLVRRHPQNGARMLLDC-RQVSALV 240
VC_1295 297 SRDNETGHSILRTQVSYVKAALAEYLAA9PAYIEILLYKSAPDHDVGKVGIPDSVLLKPGKLTDEFTVMKEHPRIQAQALIAIA6SSFLAIAK 399
VC_1348 231 TRDPETGNHLLRTQHYVVKVLAQRLA9PTVIDTYFKAAPLHDIKGVGIPDNILLKPGKLTDPDFTTMRNHALLGKLALEKA7TALINVAK 334
VCA0210 342 ARSRETGAHVQRVALYCERLAHLYGLAESADMIKNASPDHDVGKVAIPDSVLLKPGKLNDAQEWALIMQKHVEYGYEILNRSKRRLMQVAK 431
VCA0681 248 AKSQFTFQHSQKVLAELCQHLAKELGLNVEMQKALYLTGLVHDICKLHTPEELHKKPGKLNESYLCIQRHSTDYRYTLQMV--FGQSVVC 335
PF01966 1 ----RLEHSLRVALARRLAEEELG--ELDRELLLLAALLHDICKGPGFD--LEKE-----EFEILGHAVDGAETILRELKLGLEDVL 74
PF13487 1 ----SDYHVFEHSLRVAQLAAALAEELGLLDIELLLSHVAGLHDMGKIAIPDSVLLKPGKLTDEEMNVMKRHPRIGYELLSGS-ERFIQVGA 88
SM00471 1 ---EHRFEHSLRVAQLARRLAEEELGLSEEDIELLLRLAALLHDICKPGTDPAT-----TEESELEKDHAIVGAETILREL6KLIDELIL 85
cd00077 145 SRDNETGHSILRTQVSYVKAALAEYLAA9PAYIEILLYKSAPDHDVGKVGIPDSVLLKPGKLTDEFTVMKEHPRIQAQALIAIA6SSFLAIAK 232
COG2206 181 VRDYETGDLERVAQYSELLAELLGLSEEEVDLIKKAAPLHDIKGVGIPDSVLLKPGKLTSEFEIMKGPILGAETILKSS-PELMQVAA 269
COG3437 13 MKAPGTYHSMVMANLAECADKIGA---NSLLVRVGCFFHDICKTLRPPEYVENQ9LTPEQSRDILSLHFKDGAETILKEN--HMPQPII 105
4S1B 45 KSGEPYIVHPIQVAGLADLHLDA-----VTVACGFLHDV-----VEDTDITLDNIEFDFGKDVR 99
1VJ7

```

B

```

Ha4HHTT*-TT---SSS---GGG-HHHHHHα5HHHHHH*-*-*-HHHHHα6HHH---*-HHHHHα7HHHH-HHHα8HHH-
Aq_2027 102 EVLLYHHEKYDGTGYBFGKKGEETPLARLYTIAASFAMTTRRIYKRAKSWEALKELEELAGKQFDDPDIVPYAVRALSRIRALPLGEN 191
3TM8 232 NIIIGHEETINGTCFKGLREKMDPLAVLVSSANAMDRLITFEGVPK---AEAAKKLMIDHVCKH--PLQHIQHNDILKGL----- 308
4MCW 271 DIAIEHHERWDGTGYBFGKKGEETSYGRMTSISDVFDALTSDRPYKKAWMDMRTVRFFKEQKKGKHFDPFLTDIFLKNIDQMSIKREL 360
4R8Z 126 QIAYSHHERWDGRCFPEGLAGERIPLAARIVVALADRYDELTSRHAYRPPLAHAAEAVLLIQAGAGSEFDPRLVEAFVAVADAFEAARRYA 215
5Z7C 377 RMAGAHHEKLDGTGYBGRVNGDEISLMTRIITTAIDFDALSAERPYRAAMPIDKALAIMEENLHTAITPECFAALKKALNLPDEYT--- 463
BB_0374 219 LTLTHHEKNDGTGYBKGLTSENISTESNIIGAASAYSAILLDKAYKKSFSNGASIIELIKDADKKEDKRVKLKLIINAI2CPDLDIVELN 309
XC_2335 281 LIALRHHERYDGSGYBDGLVGEAIPLEARIVAVADVFDALLSARPYKEAWTMDAALAYLYAQRGRFLDPRCVDALLRGRAQEIQICGQFS 370
TM_0186 255 EIALTHHERWDGSGYBRGLKEREIPISGLIVAVADSEFAMVSRRPYKPKPLEEAFREIESLSGKLYSPEVVEAFLKLEKETDVYRREK 344
PA_4108 241 DVCLHHHERIDGTGYBFGLAQEQISLLARMGAVCVVDYDAITSDRPYKKGWNAEEAIRMAEWNG-HFDPQVFRFVAVGVIYPVGLVRL 329
VC_1295 400 EIALTHHEKWDGTGYBAQLQGEAIPLSGRMLALADVYDALISARVYKPAFSDHAKAIIVEGSGHEDPAVVEAFLAVEEKEFAIAAHFK 489
VC_1348 335 EIAMGHHEKWDGSGYBLGLKGDDIPLSARLMALADVYDALICRVYKEPMSHEEAKAIIQLGRGSHFDPMDVDAFLIEEQNFIDIAQKFA 424
VCA0210 432 EIAATHHEKWDGSGYBNRLQGDDIPLSGRIITAIADVFDALGAKRSYKDPWTDEQIREELMAQKGRHEFEKLVLELLLHWDFAIRASLP 521
VCA0681 336 EWAGNHHERLDGSGYBRGLQGAAIDPLSRIITAIADVQFALTAQRPYRGSMLSNEVMNIMRHEV2GRIDSQVFDVIVRNSQQYQLSIAES 426
PF01966 75 KLILYHHESEWAGYBGLYGEPILEARIVKLDRLDAL----- 110
PF13487 34 EIIIGQHHERLDGSGYBRGLKGEETPLGARIL----- 64
SM00471 90 LIALRHHERYDGSGYBDGLVGEAIPLEARIVAVADVFDALLSARPYKEA----- 138
cd00077 86 AVDASHHERLDGSGYBDGLKGEETLEARIVKLDRLDALRRDSREKRRRIAEEDLEELL----- 145
COG2206 233 AVALRHHERWDGTGYBRGLKGEETLEARIVAVADVYDALTSDRPYKKAKSPEEALEELRKNSGGKEDPKVVDVAFKALSKYPIGSLVLL 322
COG3437 270 EIAARHHERWDGSGYBDGLKGEETPLSARIVAIADVFDALVSGRPYKEAMSTEEALEIIRAQSGRIFDPKLVFAFQVEDEIDDIARRFA 359
4S1B 107 DIALQHGHGTLLKYFYFKAKETN17KEIAIINISDSVEAAVRSSTETPM2ITEIIDGIKDRFLDG12KIIRDTLIATLNGIYHQRIQYP 221
1VJ7 100 DIVDGVTKLGKVEYKSHEEQLAEL12IRVILVKLADRLHNMRTLKHLRKD2ERISRETMEIYAPLAH11EDLAFRYINETEFYKISHMNE 209

```

Figure 1. Structure-based sequence alignment and sequence logo of the HD-GYP domain.

A. The alignment shows the sequences of the *Aquifex aeolicus* protein Aq_2027, used in the initial description of the HD-GYP domain (3), several experimentally characterized HD-GYP domains, and HD-GYP domain entries in several public databases. Amino acids of the eponymous HD and GYP motifs are shown in white on red background. Other metal-binding residues are in white on blue background. C-di-GMP binding residues of PmGH are shown in blue and indicated by asterisks on the top line. Conserved hydrophobic residues are shaded yellow, conserved turn residues (Pro, Gly, Ala, Ser) are shaded green. The first line shows the consensus secondary structure of the domain (H, α -helix; G, 3_{10} -helix; T, turn; S, bend). The proteins are listed under their PDB IDs or genomic locus tags; their GenBank accession numbers and respective publications are listed in Table 3. For comparison, two bottom lines show the sequences of structurally characterized ‘classical’ HD (not HD-GYP) domains from two signal transduction proteins: c-di-AMP phosphodiesterase PghP (PDB: 4S1B, [CAC99544](#)) (89) and (p)ppGpp synthase/hydrolase RelA (PDB: 1VJ7, [CAA51353](#)) (102). **B.** Sequence logo of the HD-GYP domain, derived from an alignment of 1080 sequences of the RpfG (REC-HD-GYP) family.

Figure 2. Organization of the active site of the HD-GYP domain.

Amino acid residues involved in maintaining the trinuclear metal site (A) and c-di-GMP binding (B) in *Persephonella marina* protein PmGH [PDB ID: 4MCW (A) and 4MDZ (B) (45)]. Residue and guanine base numbering as in Figure 1 and Table 1. Carbon atoms of positively charged residues (His, Arg, Lys) are in blue, those of acidic residues (Asp, Glu) are in pink, and those of uncharged residues are in gray; nitrogen atoms are in dark blue, oxygens are in red. In A, Fe^{2+} ions are shown in dark red, metal coordination bonds are shown as dashed cyan lines. In B, carbon atoms of c-di-GMP are in green, phosphorus atoms are in orange, and oxygens are in red. Hydrogen bonds are shown as dashed orange lines, hydrophobic stacking is indicated by solid yellow lines.

Figure 3. HD-GYP domain content of various bacteria.

The numbers of HD-GYP-containing proteins encoded in selected bacterial groups as listed in the COG database (52). The symbols represent members of the following taxa: blue squares, *Betaproteobacteria*; green squares, *Deltaproteobacteria*; red circles, *Clostridia*; yellow circles, *Thermotogae*; dark brown circles, *Spirochaetes*; empty circles, other bacteria.