GGDEF Domain Is Homologous to Adenylyl Cyclase

Jimin Pei[2] and Nick V. Grishin[1,2]*
[1]*Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas*
[2]*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas*

ABSTRACT **The GGDEF domain is detected in many prokaryotic proteins, most of which are of unknown function. Several bacteria carry 12–22 different GGDEF homologues in their genomes. Conducting extensive profile-based searches, we detect statistically supported sequence similarity between GGDEF domain and adenylyl cyclase catalytic domain. From this homology, we deduce that the prokaryotic GGDEF domain is a regulatory enzyme involved in nucleotide cyclization, with the fold similar to that of the eukaryotic cyclase catalytic domain. This prediction correlates with the functional information available on two GGDEF-containing proteins, namely diguanylate cyclase and phosphodiesterase A of *Acetobacter xylinum*, both of which regulate the turnover of cyclic diguanosine monophosphate. Domain architecture analysis shows that GGDEF is typically present in multidomain proteins containing regulatory domains of signaling pathways or protein–protein interaction modules. Evolutionary tree analysis indicates that GGDEF/cyclase superfamily forms a large diversified cluster of orthologous proteins present in bacteria, archaea, and eukaryotes. Proteins 2001; 42:210–216.** © 2000 Wiley-Liss, Inc.

Key words: structure prediction; diguanylate cyclase; bacterial signaling pathways; ferredoxin fold; domain architecture

Sequence-based structure predictions are used routinely in structure-functional studies of proteins. Recent improvements in the detection of remote homologues by profile analysis result in the most reliable predictions. Programs such as PSI-BLAST,[1] HMMer,[2,3] and GenTHREADER,[4] combined with exponentially increasing number of sequences and determined protein structures, offer high reliability fold predictions for 30–50% of globular proteins in organisms with completely sequenced genomes.[5] These predictions can be carried out completely automatically and on a large scale.[2,4–7] However, detailed comparative study that combines the existing knowledge on a particular protein family with the general methods of protein sequence and structure analyses could extend the boundaries of entirely automated methods. This article presents an application of such an approach to deduce that GGDEF domain is homologous to adenylyl/guanylyl cyclase catalytic domain with known structure[8,9] and thus possesses the same fold. Our analysis facilitates understanding of the function and catalytic mechanism of the GGDEF domain in the absence of available three-dimensional structure.

The GGDEF domain is widespread in prokaryotes. In COG database at NCBI,[10–12] it corresponds to one of the largest clusters of potential orthologues without known or predicted spatial structure (COG2199, 70 proteins). GGDEF was first characterized as a novel domain in the response regulator PleD in *Caulobacter crescentus*, required for the swarmer-to-stalked-cell transition.[13,14] The name of this domain is due to the conserved GG[DE][DE]F sequence pattern. In *Acetobacter xylinum*, the proteins containing this domain were experimentally shown to possess diguanylate cyclase (DGC) and phosphodiesterase A (PDEA) activity.[15] These two enzymes regulate the synthesis of cyclic di-GMP, which serves as a specific nucleotide regulator of β-1,4-glucan (cellulose) synthase.[15] In both SMART[16,17] and Pfam[18] databases, GGDEF domain is characterized as a "domain of unknown function." The COG database at NCBI[10–12] annotates GGDEF according to the functional studies as diguanylate cyclase/phosphodiesterase domain 1.

We used the PSI-BLAST program[1,19] to search for homologues of the GGDEF family. PSI-BLAST searches with the default parameters (BLOSUM62 matrix,[20] 0.001 as an *E*-value threshold, no low complexity filtering in the query sequence) on the filtered for low complexity regions[21,22] nonredundant protein database (nr, March 28, 2000; 461,791 sequences) were iterated to convergence with a single copy of GGDEF domain from *Aquifex aeolicus* (NCBI gene identification (gi) number gi|2982975, residues 171–338). Found homologues were grouped by single-linkage clustering (1 bit per site threshold, about 50% identity) as implemented in SEALS package,[23] and the representative sequences were used as new queries for subsequent PSI-BLAST iterations. After three rounds of these extensive searches, about 150 GGDEF domains were detected. Surprisingly, in addition to the GGDEF domains, two close homologues of adenylyl cyclase catalytic domain (*Mycobacterium leprae* gi|3097240, residues 309–438; *Mycobacterium tuberculosis* gi|2105049, residues 328–457) were also found during the course of iterations, suggesting

Abbreviations: PDB, protein databank; DGC, diguanylate cyclase; PDEA, phosphodiesterase A.

*Correspondence to: Nick V. Grishin, HHMI, Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050. E-mail: grishin@chop.swmed.edu

Received 12 June 2000; Accepted 22 September 2000

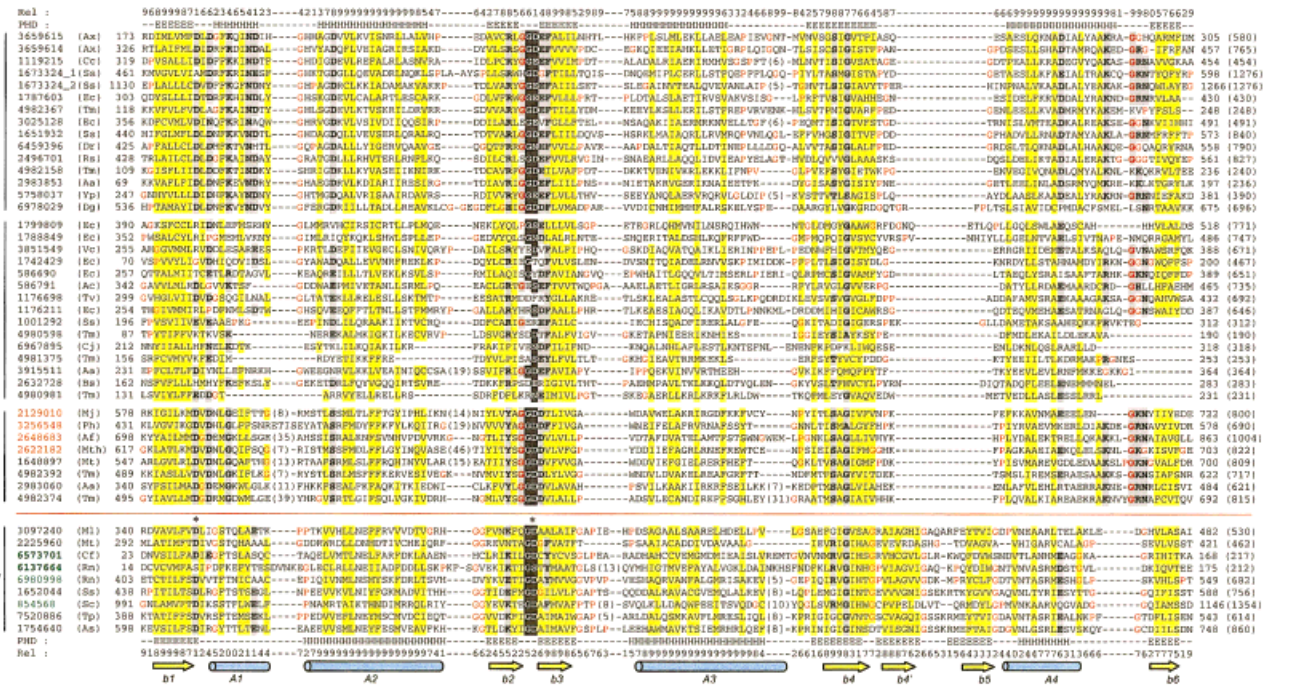

Fig. 1. Multiple sequence alignment of GGDEF/cyclase domain superfamily. Four groups of sequences (marked on the left) are shown. GGDEF family sequences are above the red line. Adenylyl/guanylyl cyclases (group IV) are below the red line. Group I represents typical GGDEF domains. Group II consists of more divergent GGDEF sequences. Group III includes archaeal GGDEF domains. Each sequence is identified by the NCBI gene identification (gi) number, followed by an abbreviation of the species name. The archaeal and eukaryotic domains are marked with red and green gi numbers, respectively, and the gi values of domains with known structure are underlined (gi|6573701, C1 domain; gi|6137664, C2 domain; PDB: 1cjv). The first and the last residue numbers of the sequences are indicated, and the total length of the proteins is shown in parenthesis at the end. Long insertions in the loop regions are not displayed but with the number of omitted residues shown in parentheses. Uncharged residues at mainly hydrophobic sites are shaded yellow. Two conserved residues in the GGDEF loop are highlighted as white bold letters on black background. Other conserved residues are shaded gray. Glycines and prolines in loop regions are shown as red letters. The two catalytically important aspartates are labeled with asterisks. The diagram of the secondary structure elements is shown at the bottom according to the structure of mammalian adenylyl cyclase catalytic domain (PDB: 1cjv, chain A). β-strands and α-helices are shown as arrows and cylinders, respectively. PHD secondary structure predictions and reliabilities (9, highest; 0, lowest) made from the GGDEF domain sequences (above the red line) and cyclase catalytic domains (below the red line) are shown on the top and at the bottom, respectively. Species names: Aa, *Aquifex aeolicus*; Ac, *Azorhizobium caulinodans*; Af, *Archaeoglobus fulgidus*; As, *Anabaena* sp.; Ax, *Acetobacter xylinus*; Bs, *Bacillus subtilis*; Cc, *Caulobacter crescentus*; Cf, *Canis familiaris*; Cj, *Campylobacter jejuni*; Dg, *Desulfovibrio gigas*; Dr, *Deinococcus radiodurans*; Ec, *Escherichia coli*; Mj, *Methanococcus jannaschii*; Ml, *Mycobacterium leprae*; Mt, *Mycobacterium tuberculosis*; Mth, *Methanobacterium thermoautotrophicum*; Ph, *Pyrococcus horikoshii*; Rn, *Rattus norvegicus*; Rs, *Rhizobium* sp.; Sc, *Saccharomyces cerevisiae*; Ss, *Synechocystis* sp.; Tm, *Thermotoga maritima*; Tp, *Treponema pallidum*; Tv, *Thiocystis violacea*; Vc, *Vibrio cholerae*; Yp, *Yersinia pestis*.

possible homology between the GGDEF and cyclase catalytic domains. Indeed, PSI-BLAST search initiated with the GGDEF domain from *Streptomyces coelicolor* (gi|6580642, residues 268–438) detected the two *Mycobacterium* cyclase domains with *E* values 0.002 and 0.003, respectively, on the second iteration (May 23, 2000; 504,523 sequences in nonredundant database; default parameters). As the spatial structure of adenylyl cyclase catalytic domain is available,[8,9] this weak but detectable and statistically supported sequence similarity might offer the fold prediction for GGDEF domain.

To probe the PSI-BLAST-based prediction further, we applied several fold recognition (threading) methods to the representatives of GGDEF family. The following methods were explored: (1) the hybrid fold recognition method of Fischer,[24] as implemented at the BIOINBGU server (http://www.cs.bgu.ac.il/~bioinbgu/); (2) the GenTHREADER program[4] at the PSIPRED server (http://insulin.brunel.ac.uk/psipred/); (3) the secondary structure prediction-based threading TOPITS[25,26] at http://www.embl-heidelberg.de/predictprotein/; and (4) the method of Fischer and Eisenberg[27] and Rice and Eisenberg[28] at the University of California (UCLA) fold recognition server (http://fold.doe-mbi.ucla.edu/). The consensus method of Fischer, which combines sequence, structural, and evolutionary information,[24] did give the cyclase fold among the top hits for many GGDEF family sequences. For example, one of the GGDEF domains from *Synechocystis* sp. (gi|1651932, 440–573; Fig. 1) found adenylyl or guanylyl cyclases as the top four hits with consensus scores ranging from 5.5 to 8.6 (1awn,[29] chain B: 8.6; 1awm,[29] chain A: 8.5; 1ab8,[8] chain A: 6.6; 1azs,[9] chain A: 5.5). The other three threading methods did not yield significant fold recognition results. However, for many GGDEF sequences, they found hits to the ferrodoxin fold[30,31] that adenylyl or guanylyl cyclases possess. For example, GenTHREADER with the GGDEF domain from *Thermotoga maritima* (gi|4982367, 118–248; Fig. 1) as a query found the C-terminal domains of

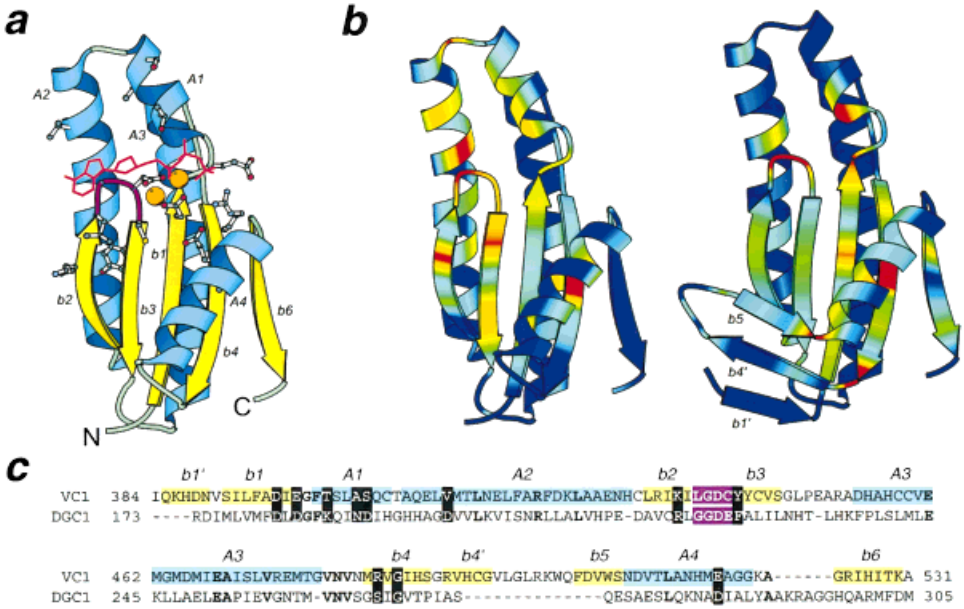

Fig. 2. Structural diagrams of GGDEF/cyclase domain. **a:** Mammalian adenylyl cyclase catalytic domain (PDB: 1cjv, chain A, residues 384–531) showing the secondary structure elements present in both cyclase catalytic domain and GGDEF domain. Secondary structure elements are labeled by italicized letters and numbers. α-helices and β-strands are labeled with letters A and b, respectively, followed by the corresponding number. GGDEF loop is shown in magenta. Conserved residues in GGDEF domain are shown in ball-and-stick representations of their counterparts in the cyclase catalytic domain. ATP is displayed in red lines; two metal ions are shown as orange balls. Drawn with the program BOBSCRIPT.[58] **b** Ribbon diagrams showing sequence conservation in multiple alignments of GGDEF domain family (left) and cyclase catalytic domain (right). Red and blue correspond to highest and the lowest conservation, respectively. **c:** Sequence alignment of mammalian catalytic domain (VC1; PDB: 1cjv, chain A) and GGDEF domain of diguanylate cyclase (DGC1) of *Acetobacter xylinus*. Residue numbers are indicated for the first amino acid residue in each line and for the last amino acid residue in each sequence. The labeling and color shading of secondary structure elements correspond to **a**. Identical residues in the two sequences are shown in bold letters. Conserved residues of GGDEF family and their counterparts in cyclase domain are shown as white letters on black background.

arginine repressor from *Bacillus stearothermophilus* (1b4b chain A[32]) and *Escherichia coli* (1xxa chain A[33]) as the first two hits with the marginal significance (probability of correct match 0.3). Arginine repressors possess a circularly permuted ferredoxin-like fold (ββαββα in arginine repressor versus βαββαβ in ferredoxin fold) and thus are structurally similar to adenylyl/guanylyl cyclases. GenTHREADER aligned βαββα unit of arginine repressor to the GGDEF sequence placing the GGDEF-containing loop between the two β-strands structurally equivalent to the corresponding β-strands in adenylyl/guanylyl cyclases.

To validate the hypothesized homology, multiple sequence alignments of both GGDEF and cyclase catalytic domain families were constructed using the ClustalW program.[34] Global alignment was adjusted manually based on PSI-BLAST local alignments to match the conserved motifs (Fig. 1). Secondary structure predictions for both families were carried out using PHD server.[35] The secondary structure predictions show an excellent correspondence between the GGDEF and cyclase families, as well as between the predicted and experimentally determined structures of the cyclase catalytic domains. (Fig. 1). The cyclase domain adopts a ferredoxin-like fold with βαββαβ secondary structural pattern corresponding to b1, A2, b2, b3, A3, and b4 of the domain core (Figs. 1, 2). Ferredoxin fold $(\beta\alpha\beta)_2$ is characterized by a duplication of the right-handed split βαβ unit. The connection αββ is left-handed. This fold represents the most commonly encountered arrangement of the anti-parallel β-sheet flanked by α-helices. It has been shown that the "palm" domain of *Pol* I family polymerases[36,37] is homologous to cyclase catalytic domain.[38,39] Cyclases and polymerases display very similar active site arrangements and catalytic mechanisms. In contrast to other ferredoxin fold proteins, both cyclase and polymerase domains have longer α-helices (A2 and A3) and an additional short α-helix (A1) after the first β-strand (b1) (Fig. 2). Both the GGDEF and cyclase catalytic domains possess a sequence segment corresponding to the short insertion helix A1, which is predicted to adopt helical conformation in GGDEF domain (Fig. 1). In cyclases, a ferredoxin-like core is extended by an α-helix (A4) and a β-strand (b6) at the C-terminus. The predicted C-terminal α-helical segment (A4) is present in all GGDEF domains, and the sequence corresponding to b6 is found in most of these domains.

Besides the conservation of the predicted secondary structure elements, which is also reflected in the specific patterns of hydrophobicity (Fig. 1, yellow shading), several motifs that involve conserved charged residues span the entire length of the GGDEF/cyclase alignment. These include functionally and structurally important parts of the sequence. First, two Mg^{2+}-binding aspartate residues of cyclase domain, one at the end of b1 and another between b2 and b3 (Asp396 and Asp440; Fig. 2c), are conserved in the GGDEF domain. The aspartate between b2 and b3 maps to the GGDEF loop (Figs. 1, 2). Second, the

glycine in the middle of b4, a highly conserved small residue in cyclase catalytic domain due to the steric restriction caused by the packing of A4 against b4 (Fig. 2a), is also conserved in GGDEF domain (Figs. 1, 2). Based on all these common features, such as conservation of secondary structure predictions, hydrophobicity patterns, and functionally and structurally important residues clustered in conserved motifs, we propose that the two families are evolutionarily related and unify them to GGDEF/cyclase superfamily.

The alignment also demonstrates several structural differences between GGDEF and cyclase families. The two separate subdomains of the cyclase catalytic domain[8,9] have no counterparts in the GGDEF domain. One is placed at the C-terminus of cyclase catalytic domain (after β-strand b6, not shown in Fig. 2). The other subdomain is composed of the three antiparallel β-strands formed by the extension of b1 (b1′ in Fig. 2b) at the N-terminus, and the loop between b4 and A4 structured as a β-hairpin (b4′ and b5, Fig. 2b). The GGDEF domain lacks the insertion between b4 and A4 (Fig. 1); however, it possesses an additional conserved N-terminal sequence fragment that is predicted to form two α-helices, and not β-strands (data not shown).

Extensive sequence similarity searches also revealed many divergent copies of GGDEF domains (group II, Fig. 1). The four archaeal proteins, together with a few bacterial proteins, form another distinct group (group III, Fig. 1) with slightly different conservation pattern and longer loop regions.

The distinctive feature of the GGDEF/cyclase superfamily as well as of the DNA polymerase "palm" domains is the conservation of catalytically important aspartate residues. These two residues (marked by an asterisk, Fig. 1), one at the end of b1 and the other in the loop between b2 and b3, are essential for coordinating the two metal ions in the catalytic core domain[8,9,39,40] (Fig. 2a). They contribute to the two-metal-ion catalytic mechanism shared also by many other enzymes.[40] Based on this conservation, we infer that GGDEFs are likely to be the enzymatic domains in the DGC and PDEA that catalyze synthesis or hydrolysis of cyclic di-GMP by the two-metal-ion mechanism.

The sequence alignment shows that GGDEF domains contain larger numbers of conserved charged or polar residues than cyclase catalytic domains (Figs. 1, 2a,c). All these conserved residues are clustered around the active site in the structure (Fig. 2a,b). DGC and PDEA catalyze the linkage between the two GTP molecules, in contrast to the cyclization of one ATP/GTP molecule in adenylyl/guanylyl cyclases; thus, more substrate/cofactor binding residues might be needed in the GGDEF domains. Mammalian adenylyl cyclases function as heterodimers of two homologous catalytic domains, termed C1 and C2.[39,41] The C2 domain lacks the two metal-binding aspartate residues (Fig. 1, gi|6177664). However, it contributes several basic residues (Arg160/1029, Lys196/1065; PDB: 1cjv) around the catalytic core of the C1 domain (Fig. 1, gi|6573701) to stabilize the phosphate groups. (The first number corresponds to the crystallized fragment, and the second number to the full-length protein.) There are also a few residues (Lys69/938, Asp149/1018, pdb: 1cjv; Glu925, Cys995 in corresponding guanylyl cyclase[42]) that help define the substrate specificity.[39,42] These residues function as a hydrogen bond donor to (Lys938 and Cys995) and acceptor (Asp1018 and Glu925) from the purine ring. Lys938/Glu925 is in the middle of b2. In typical GGDEF domains (group I, Fig. 1), this position is occupied by conserved positively charged residues (in most cases, Arg) and may also contribute to the specificity of GGDEF domains. However, the Asp1018/Cys995 is located in the insertion between b4 and A4, which is absent in GGDEF members. There is no evidence yet that GGDEF domains form dimers. If they form heterodimers like the mammalian adenylyl cyclases, the domains of the group II (Fig. 1) might function as C2 domains since the Mg-binding aspartates are not present in many of them, and typical GGDEF domains (group I) should be functional homologues of C1 domain. Alternatively, GGDEF cyclases might function as homodimers with symmetric active sites, as occurs with some guanylyl cyclases.[29,43,44]

We compared conservation of sites between GGDEF and cyclase domains. The spatial distribution of sequence conservation in the structures of both domains has largely similar properties (Fig. 2b): the β-sheet in the middle is more conserved than the surrounding α-helices, and residues around the catalytic core are more conserved than other parts of the molecule. The striking conservation of a glycine residue in the middle of the β-strand b4 indicates steric restrictions caused by packing of A4. Several differences in the conservation are noticeable. In addition to the conservation of the loop between b2 and b3, GGDEF domains carry highly conserved sites in the flanking β-strands b2 and b3. Conserved glycine residues cap the β-hairpin b4′-b5 in cyclases but are lacking in GGDEF domain due to the absence of the hairpin.

To clarify the evolutionary relationship between GGDEF and cyclase catalytic domains, an evolutionary tree from the representative sequences of both families was constructed using MLd2tree (Y.I. Wolf and N.V. Grishin, unpublished) and PHYLIP[45] programs (Fig. 3). As many sequences of the GGDEF/cyclase superfamily are very divergent and the domain is relatively short (about 150 residues), the estimates of evolutionary distances between the sequences are characterized by large standard errors and the tree branching cannot be resolved with confidence. Thus, the tree shown represents a classification diagram for the superfamily rather than a reliable scheme of phylogenetic events. However, the unrooted tree clearly shows that cyclase domains (group IV, Fig. 1) and archaeal group (group III, Fig. 1) of GGDEF domains form separate clusters. We term these eukaryotic and archaeal clusters, respectively, implying that bacterial members in them represent horizontal gene transfer events. The typical bacterial GGDEF domains (group I, Fig. 1) cluster together with the divergent bacterial GGDEF domains (group II, Fig. 1) to form a bacterial group. This group is characterized by highly uneven branch lengths resulting from the variability of substitution rates between proteins.

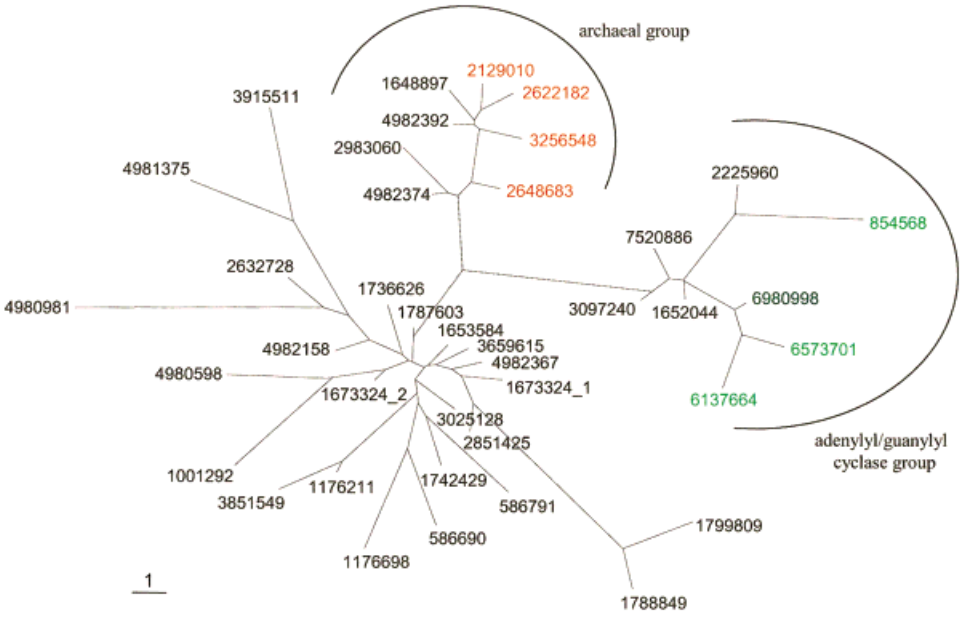

Fig. 3. Phylogenetic tree of GGDEF/cyclase domain superfamily. Sequences are labeled with gene identification (gi) numbers explained in the legend to Figure 1. Adenylyl/guanylyl cyclase group and the archaeal group of GGDEF domains are outlined. Archaeal and eukaryotic sequences are labeled in red and green numbers, respectively. Scale bar corresponds to evolutionary distance of 1 amino acid substitution per site. The tree was constructed with the MLd2tree program (Y.I. Wolf and N.V. Grishin, unpublished), which took into account substitution rate difference among sites for distance calculation and used neighbor-joining method, as implemented in PHYLIP package.[45]

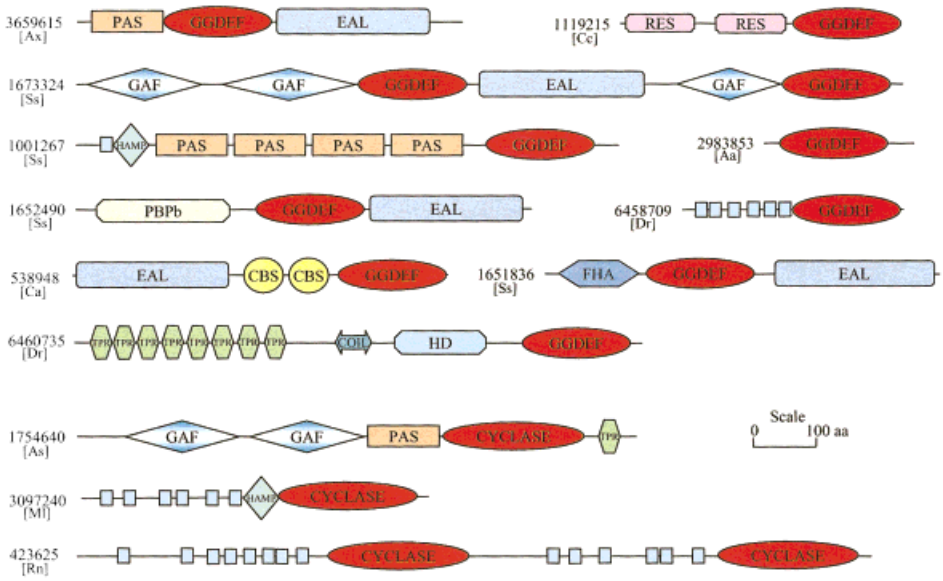

Fig. 4. Domain architecture of selected GGDEF/cyclase superfamily proteins. Predicted transmembrane regions are shown as small blue rectangular boxes. The length of each sequence is approximately to scale. The proteins are labeled with gi numbers and abbreviations of species names according to Figure 1: 3659615, diguanylate cyclase [*Acetobacter xylinus*]; 1119215, pleD gene product [*Caulobacter crescentus*]; 1673324, hypothetical protein [*Synechocystis* sp.]; 1001267, PleD [*Synechocystis* sp.]; 2983853, hypothetical protein [*Aquifex aeolicus*]; 1652490, hypothetical protein [*Synechocystis* sp.]; 6458709, GGDEF family protein [*Deinococcus radiodurans*]; 539848, hypothetical protein, [*Clostridium acetobutylicum*] (in nonredundant database, this protein appears as a fragment truncated at the N-terminus, the full length proteins was taken from WIT site http://igweb.integratedgenomics.com/IGwit/); 1651836, hypothetical protein [*Synechocystis* sp.]; 6460735, GGDEF family protein [*Deinococcus radiodurans*]; 1754640, adenylyl cyclase cyaB2 gene product [*Anabaena* sp.]; 3097240, putative membrane protein [*Mycobacterium leprae*]; 423625, adenylyl cyclase type 5 [*Rattus norvegicus*]. Domain names: EAL, domain of unknown function containing conserved EAL motif; RES, response regulator CheY-like domain; GAF (cGMP-specific and -stimulated phosphodiesterases, *Anabaena* adenylyl cyclases and *Escherichia coli* FhlA) domain; HAMP, histidine kinases, adenylyl cyclases, methyl binding proteins, phosphatases) domain; PAS, PAS and PAC domains together; PBPb, periplasmic substrate-binding protein (bacterial) domain; CBS, cystathionine-β-synthase domain; FHA, forkhead-associated domain; TPR, tetratricopeptide repeat; HD, a phosphohydrolase domain that contains conserved HD motif. COIL is the predicted coiled coil segment.

The high variability and low conservation of the divergent GGDEF domains can be easily seen from the alignment (group II, Fig. 1). Despite this variability, divergent GGDEF domains group with the typical bacterial GGDEF domains and do not show affinity to archaeal and eukaryotic clusters (Fig. 3). The rooting of all three groups (bacterial, archaeal, and eukaryotic) is tentative due to the low sequence conservation between groups and small domain size (dotted lines, Fig. 3). The most striking feature of the tree is that the evolutionary distances between pairs of divergent GGDEF domains are not smaller (and for some, they are greater, e.g., gi|3915511 and gi|1788849) than some of the distances between eukaryotic cyclases and GGDEF domains (e.g., gi|7520886 and gi|1736626). This is merely a reflection of the fact that the sequence identity (similarity) between divergent GGDEF domains is smaller than that between the closest members of GGDEF and cyclase families. For example, identities between gi|3915511 and gi|1788849 and between gi|7520886 and gi|1736626 are 7.6% and 10%, respectively. From the tree, it seems likely that prokaryotic GGDEF domains are orthologues of eukaryotic cyclases and GGDEF/cyclase superfamily represents a single large and diversified orthologous cluster.

Overall we detected about 150 copies of GGDEF domains in nonredundant protein database (March 28, 2000: 461,791 sequences). One of the characteristic features of GGDEF domains is their presence in multiple copies (that share 10–90% identity between them) within single genomes. The largest number of GGDEF domains (in 22 proteins) was detected in *Synechocystis* sp. Other organisms with high-copy numbers of GGDEFs are *Escherichia coli* (21), *Deinococcus radiodurans* (16), *Thermotoga maritima* (13), and *Aquifex aeolicus* (12). Therefore, it is likely that the products of reactions catalyzed by these GGDEF cyclases serve as important nucleotide regulators or metabolites of biological processes in bacterial species.

Domain architecture analysis of the GGDEF domain-containing proteins was performed using SMART[16,17] and Pfam[18] databases (Fig. 4). GGDEF domain is usually present in multidomain proteins, with only a few forming single-domain proteins. About 30 GGDEF proteins contain predicted transmembrane regions and are likely to be intrinsic membrane proteins. Some proteins have signaling peptides at the N-termini and one (gi|1652490) has the periplasmic substrate-binding domain (PBPb),[16,46] suggesting that they are periplasmic proteins. About 10 different domains co-occur with GGDEF domain (Fig. 4). The most frequent domains are often regulatory domains in signaling pathways such as the PAS/PAC domain,[47] the response regulator CheY-like domain,[48] the HAMP domain,[49] and the GAF domain.[50] The EAL domain[15] (named after the conserved EAL signature motif), whose function is unknown, co-occurs with GGDEF domain particularly frequently. There are three DGCs and three PDEAs in *Acetobacter xylinum,*[15] all of which have the similar domain architecture containing PAS/PAC, GGDEF, and EAL domains. The PleD of *Caulobacter crescentus* contains one copy of GGDEF domain and 2 copies of response regulator CheY-like domains.[13] Many of these domains can mediate protein–protein or protein–ligand interactions, like TPR,[51] PAS/PAC,[47] FHA[52] (which is shown to be able to bind phosphopeptides), and GAF[50] (which can serve as regulators by binding nucleotides or small molecules). In contrast to bacterial GGDEF, mammalian adenylyl cyclases are usually membrane proteins containing two tandem copies of catalytic domains[41] (C1 and C2) (Fig. 4). The bacterial homologues of these cyclases often co-occur with such domains as HAMP and GAF, similarly to GGDEF domains. These additional domains can serve as regulators of the catalytic functions of GGDEF/cyclase domains in response to changes of environmental factors or in signaling pathways.

The potential products of GGDEF domains include cyclic di-GMP (by DGCs) or its hydrolytic product (by PDEAs).[15] Cyclic di-GMP is known to be important for activation of cellulose synthase in *Acetobacter xylinum*.[53,54] The involvement of GGDEF domain in cellulose synthesis has also been demonstrated for *Rhizobium leguminosarum*.[55] This regulation may also be used by other bacteria. Loss of PleD gene function in *Caulobacter cresentus* results in prevention of stalk formation during cell cycle.[13,14] Mutagenesis demonstrated that GGDEF domain was essential for this phenotype.[14] Stalk formation may need synthesis of polysaccharides, which requires enzymes like cellulose synthase.[56,57] We suggest that PleD regulates the level of cyclic di-GMP, which may activate enzymes participating in stalk formation. Whether the GGDEF domain can catalyze the cyclization of other nucleotides remains unknown.

In summary, through the sequence and structure analysis of GGDEF domains as well as cyclase catalytic core domains, we demonstrate homology between these proteins and predict the fold of the GGDEF domain. This homology implies the functional prediction for the GGDEF domain as well as the prediction of its catalytic mechanism. GGDEF domain is likely to possess nucleotide cyclase activity, such as cyclic di-GMP formation, with a mechanism similar to that employed by the adenylyl cyclase and DNA-polymerase "palm" domain.

ACKNOWLEDGMENTS

The authors are grateful to Hong Zhang and Joseph Gao for critical reading of the manuscript and helpful comments.

REFERENCES

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
2. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
3. Eddy SR. Hidden Markov models. Curr Opin Struct Biol 1996;6: 361–365.
4. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
5. Wolf YI, Brenner SE, Bash PA, Koonin EV. Distribution of protein folds in the three superkingdoms of life. Genome Res 1999;9:17–26.

6. Gerstein M, Levitt M. A structural census of the current population of protein sequences. Proc Natl Acad Sci USA 1997;94:11911–11916.
7. Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J Mol Biol 1997;274:562–576.
8. Zhang G, Liu Y, Ruoho AE, Hurley JH. Structure of the adenylyl cyclase catalytic core. Nature 1997;386:247–253.
9. Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR. Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gsalpha.GTPgammaS. Science 1997;278:1907–1916.
10. Koonin EV, Tatusov RL, Galperin MY. Beyond complete genomes: from sequence to structure and function. Curr Opin Struct Biol 1998;8:355–363.
11. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–637.
12. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28:33–36.
13. Hecht GB, Newton A. Identification of a novel response regulator required for the swarmer- to-stalked-cell transition in *Caulobacter crescentus*. J Bacteriol 1995;177:6223–6229.
14. Aldridge P, Jenal U. Cell cycle-dependent degradation of a flagellar motor component requires a novel-type response regulator. Mol Microbiol 1999;32:379–391.
15. Tal R, Wong HC, Calhoon R, Gelfand D, Fear AL, Volman G, Mayer R, Ross P, Amikam D, Weinhouse H, Cohen A, Sapir S, Ohana P, Benziman M. Three cdg operons control cellular turnover of cyclic di-GMP in *Acetobacter xylinum*: genetic organization and occurrence of conserved domains in isoenzymes. J Bacteriol 1998;180:4416–4425.
16. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 2000;28:231–234.
17. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA 1998;95:5857–5864.
18. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2000;28:263–266.
19. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998;23:444–447.
20. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
21. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18:269–285.
22. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
23. Walker DR, Koonin EV. SEALS: a system for easy analysis of lots of sequences. Intell Syst Mol Biol 1997;5:333–339.
24. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pacific Symp Biocomputing 2000:119–130.
25. Rost B. TOPITS: threading one-dimensional predictions into three-dimensional structures. Ismb 1995;3:314–321.
26. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.
27. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
28. Rice DW, Eisenberg D. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
29. Liu Y, Ruoho AE, Rao VD, Hurley JH. Catalytic mechanism of the adenylyl and guanylyl cyclases: modeling and mutational analysis. Proc Natl Acad Sci USA 1997;94:13414–13419.
30. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
31. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
32. Ni J, Sakanyan V, Charlier D, Glansdorff N, Van Duyne GD. Structure of the arginine repressor from *Bacillus stearothermophilus*. Nature Struct Biol 1999;6:427–432.
33. Van Duyne GD, Ghosh G, Maas WK, Sigler PB. Structure of the oligomerization and L-arginine binding domain of the arginine repressor of *Escherichia coli*. J Mol Biol 1996;256:377–391.
34. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
35. Rost B, Sander C, Schneider R. PHD—an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994;10:53–60.
36. Doublie S, Tabor S, Long AM, Richardson CC, Ellenberger T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. Nature 1998;391:251–258.
37. Doublie S, Ellenberger T. The mechanism of action of T7 DNA polymerase. Curr Opin Struct Biol 1998;8:704–712.
38. Artymiuk PJ, Poirrette AR, Rice DW, Willett P. A polymerase I palm in adenylyl cyclase? Nature 1997;388:33–34.
39. Tesmer JJ, Sprang SR. The structure, catalytic mechanism and regulation of adenylyl cyclase. Curr Opin Struct Biol 1998;8:713–719.
40. Tesmer JJ, Sunahara RK, Johnson RA, Gosselin G, Gilman AG, Sprang SR. Two-metal-ion catalysis in adenylyl cyclase. Science 1999;285:756–760.
41. Sunahara RK, Dessauer CW, Gilman AG. Complexity and diversity of mammalian adenylyl cyclases. Annu Rev Pharmacol Toxicol 1996;36:461–480.
42. Tucker CL, Hurley JH, Miller TR, Hurley JB. Two amino acid substitutions convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. Proc Natl Acad Sci USA 1998;95:5993–5997.
43. Yang RB, Garbers DL. Two eye guanylyl cyclases are expressed in the same photoreceptor cells and form homomers in preference to heteromers. J Biol Chem 1997;272:13738–13742.
44. Wilson EM, Chinkers M. Identification of sequences mediating guanylyl cyclase dimerization. Biochemistry 1995;34:4696–4701.
45. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996;266:418–427.
46. Hsiao CD, Sun YJ, Rose J, Wang BC. The crystal structure of glutamine-binding protein from *Escherichia coli*. J Mol Biol 1996;262:225–242.
47. Zhulin IB, Taylor BL, Dixon R. PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox. Trends Biochem Sci 1997;22:331–333.
48. Pao GM, Saier MH Jr. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J Mol Evol 1995;40:136–154.
49. Aravind L, Ponting CP. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. FEMS Microbiol Lett 1999;176:111–116.
50. Aravind L, Ponting CP. The GAF domain: an evolutionary link between diverse phototransducing proteins. Trends Biochem Sci 1997;22:458–459.
51. Sikorski RS, Boguski MS, Goebl M, Hieter P. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. Cell 1990;60:307–317.
52. Durocher D, Henckel J, Fersht AR, Jackson SP. The FHA domain is a modular phosphopeptide recognition motif. Mol Cell 1999;4:387–394.
53. Ross P, Mayer R, Weinhouse H, Amikam D, Huggirat Y, Benziman M, de Vroom E, Fidder A, de Paus P, Sliedregt LA. The cyclic diguanylic acid regulatory system of cellulose synthesis in *Acetobacter xylinum*. Chemical synthesis and biological activity of cyclic nucleotide dimer, trimer, and phosphothioate derivatives. J Biol Chem 1990;265:18933–18943.
54. Ross P, Mayer R, Benziman M. Cellulose biosynthesis and function in bacteria. Microbiol Rev 1991;55:35–58.
55. Ausmees N, Jonsson H, Hoglund S, Ljunggren H, Lindberg M. Structural and putative regulatory genes involved in cellulose synthesis in *Rhizobium leguminosarum* bv. *trifolii*. Microbiology 1999;145(pt 5):1253–1262.
56. Ong CJ, Wong ML, Smit J. Attachment of the adhesive holdfast organelle to the cellular stalk of *Caulobacter crescentus*. J Bacteriol 1990;172:1448–1456.
57. Gober JW, Marques MV. Regulation of cellular differentiation in *Caulobacter crescentus*. Microbiol Rev 1995;59:31–47.
58. Esnouf RM. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. J Mol Graph Model 1997;15:133–138.