Structural Classification of Bacterial Response Regulators: Diversity of Output Domains and Domain Combinations†

Michael Y. Galperin*

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

Received 12 December 2005/Accepted 28 March 2006

CheY-like phosphoacceptor (or receiver [REC]) domain is a common module in a variety of response regulators of the bacterial signal transduction systems. In this work, 4,610 response regulators, encoded in complete genomes of 200 bacterial and archaeal species, were identified and classified by their domain architectures. Previously uncharacterized output domains were analyzed and, in some cases, assigned to known domain families. Transcriptional regulators of the OmpR, NarL, and NtrC families were found to comprise almost 60% of all response regulators; transcriptional regulators with other DNA-binding domains (LytTR, AraC, Spo0A, Fis, YcbB, RpoE, and MerR) account for an additional 6%. The remaining one-third is represented by the stand-alone REC domain (~14%) and its combinations with a variety of enzymatic (GGDEF, EAL, HD-GYP, CheB, CheC, PP2C, and HisK), RNA-binding (ANTAR and CsrA), protein- or ligand-binding (PAS, GAF, TPR, CAP_ED, and HPt) domains, or newly described domains of unknown function. The diversity of domain architectures and the abundance of alternative domain combinations suggest that fusions between the REC domain and various output domains is a widespread evolutionary mechanism that allows bacterial cells to regulate transcription, enzyme activity, and/or protein-protein interactions in response to environmental challenges. The complete list of response regulators encoded in each of the 200 analyzed genomes is available online at http://www.ncbi.nlm.nih.gov/Complete_Genomes/RRcensus.html.

It has been 20 years since the first bacterial response regulator genes were sequenced (35, 39, 89) and the CheY-like receiver (phosphoacceptor) domain was identified as their common regulatory module (95, 124, 138; see references 55, 57, 127, and 150 for reviews and reference 43 for a historical perspective). Structural characterization of the CheY-like receiver domain (hereafter, the REC domain) from *Escherichia coli* and other organisms confirmed that it is an autonomously folding, evolutionarily stable, compact structural unit (125, 126, 147) whose conformation undergoes a distinctive change upon phosphorylation (65, 75). These conformational changes, which increase the propensity of the REC domain to form dimers, are used by bacterial and archaeal cells to transmit and propagate a wide variety of environmental and intracellular signals (29, 127, 150).

The typical scheme of bacterial two-component signal transduction involves signal sensing (ligand binding or other conformational change) by a sensory histidine kinase that leads to its autophosphorylation, followed by phosphoryl transfer to the Asp residue in the N-terminal REC domain of the cognate response regulator, which affects the properties of the C-terminal DNA-binding domain (127, 150). Structural characterization of the DNA-binding domains of various response regulators revealed several variations on the common helix-turn-helix (HTH) theme, exemplified by the NarL-type, OmpR-type "winged helix," SpolA-type, and Fis-like structures (12, 78, 79, 87, 88, 100). Sequence comparisons revealed additional types of DNA-binding domains in certain response regulators, some of which have been experimentally verified (94, 113).

In addition to forming associations with various DNA-binding domains, the REC domain can function as a stand-alone module. Its propensity to protein-protein interactions plays a key role in the chemotaxis machinery, as well as in some other regulatory pathways (10, 130). It also forms other types of response regulators by associating with certain enzymatic or protein-binding domains, e.g., with the methylesterase (CheB) and chemotaxis modulator (CheW) domains, both of which participate in the adaptation to attractants during chemotaxis (61). In the CheB-type response regulator, the N-terminal REC domain packs against the active site of the C-terminal methylesterase domain and inhibits methylesterase activity by restricting access to the active site (34). The conformational change occurring upon phosphorylation of the REC domain apparently disrupts this interaction and relieves the methylesterase domain from inhibition by REC (34). In the *Bacillus subtilis* CheV protein, the presence of the N-terminal CheW-like domain was shown to stabilize the phosphorylated state of the C-terminal REC domain (40, 61).

The REC domain can also form combinations with other signaling domains, e.g., in the *Caulobacter crescentus* PleD response regulator, which controls flagellum ejection and stalk formation during the *C. crescentus* life cycle (1, 2, 52). In PleD, the N-terminal REC domain is fused to an inactivated REC domain and a C-terminal GGDEF domain, which has diguanylate cyclase activity that produces bis-(3′→5′)-cyclic diguanosine monophosphate (c-di-GMP), a secondary messenger in bacteria (99, 112; see references 58 and 109 for reviews). In *Vibrio cholerae* response regulator VieA, which controls biofilm for-
mation by this organism (137), the REC domain is associated with the EAL domain, a c-di-GMP-specific phosphodiesterase (22, 30, 114, 131). A combination of the REC domain with another c-di-GMP-specific phosphodiesterase, the HD-GYP domain, controls biosynthesis of extracellular polysaccharide in Xanthomonas campestris (45, 111, 121).

The response regulator sets encoded in E. coli, B. subtilis, Synechocystis sp., Streptococcus pneumoniae, and several other bacteria have been tabulated and classified (48, 51, 74, 90, 91, 97, 146, 148). There have been several attempts to compile comprehensive lists of response regulators in all sequenced microbial genomes (11, 66, 70, 141), as well as to list response regulators in microbial genome databases, such as KEGG, SENTRA, and COG (60, 84, 134). However, the extreme diversity of REC domain combinations encoded in recently sequenced microbial genomes continues to generate problems during genome annotation (43, 157). We have already discussed the most common errors in annotating signaling proteins and suggested that, unless exact function of the newly described protein is known, its annotation should be based on its domain composition, rather than on sequence similarity of any given domain (38, 109). To assist in this process, I present here a comprehensive domain-based classification of the response regulators encoded in completely sequenced genomes of 200 bacterial and archaeal species and describe domain architectures that involve the REC domain and known DNA-binding, enzymatic, and ligand-binding domains, as well as several new domains, not described previously.

**MATERIALS AND METHODS**

**Data sources.** The analyzed set included complete genome sequences of 200 bacterial and archaeal species, sequenced by 20 September 2005, as listed on the NCBI’s Entrez Genomes website (151). Only one representative genome per species was chosen, typically the first publicly released one, according to the NCBI Genomes database listing. As an exception, two strains of E. coli, K12 and O157:H7, and three serovars of Salmonella enterica, Paratyphi, Typhi, and Typhimurium, were included in the list, as described earlier (42). The genome sequences were downloaded from the NCBI’s Genomes database or searched directly on the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST/) using the PSI-BLAST (3) tool.

**Identification of REC-containing proteins.** The complete list of REC-containing proteins encoded in each of the 200 analyzed genomes was compiled separately for each bacterial phylum based on the results of iterative PSI-BLAST searches (3) of the NCBI’s Reference Sequence (RefSeq) database (103). In each case, the search was initiated with the 122-amino-acid sequence of the CheY protein TM0468 (GenBank accession no. AAD35552; UniProt entry O9WYT9) and a precomputed position-specific scoring matrix. The query sequence and the position-specific scoring matrix are both available in the supplemental material. The database search space was limited to the given phylum (e.g., Firmicutes[orgn]); sequences of incomplete genomes were excluded using the srbdb_reseq_provisional[prop] modifier. The PSI-BLAST searches used strict inclusion threshold E values of 10^-4 to 10^-6, adjusting as necessary, and were repeated until convergence. The phylogenetic representation and the total numbers of REC-containing proteins encoded in each given genome were estimated using the Taxonomy Report option in the BLAST output. Potential false-positive hits were checked at every step of PSI-BLAST using the CDD Domain viewer (86) and manually removed (unselected) from the hit list for each iteration of PSI-BLAST. Low-scoring hits (potential false negatives) were inspected on a case-by-case basis by comparing them against the CDD and Pfam domain databases (14, 85) using the CDD search tool (86) with relaxed cutoff values.

**Classification of REC-containing proteins.** The lists of REC-containing proteins from each complete genome were checked against the CDD, COG, and Pfam domain databases (14, 85, 134), using the CDD Domain viewer (86), and grouped according to their domain composition. The sequences were also matched against UniProt (13), using the UniProt batch retrieval tool, and linked to the corresponding entries in the Pfam database (14). The families of response regulators were named after their experimentally characterized representatives, where available (e.g., OmpR-like and Spo0A-like). In the absence of experimentally characterized representatives, families of response regulators were named based on their domain architectures (e.g., REC-AraC and RpoE-REC). To account for the cases where different species (e.g., three species of Brucella) had merged RefSeq entries, PSI-BLAST searches were also run against selected genomes using the NCBI’s Genomic BLAST tool (31), followed by manual analysis of the outputs. Whenever possible, the protein counts were compared to the published data, and discrepancies were resolved on a case-by-case basis. In most instances, discrepancies could be attributed to different treatments of highly diverged and truncated sequences. In this work, such sequences were defined as ones showing up in the PSI-BLAST searches but not recognized as containing the REC domain by RPS-BLAST comparison against CDD, COG, or Pfam databases with default or relaxed (E value of 1.0) parameters. These were typically excluded from the total count but still listed (marked with asterisks) in Table S1 in the supplemental material. All N-terminally truncated sequences were manually inspected to see if the apparent truncation could be due to the incorrect selection of the start codon during automated annotation. Corrected open reading frames for nine genes, for which this proved to be the case (Bacillus subtilis BC5352; Bacteroides thetaiotaomicron BT4258; Bradyrhizobium japonicum bsl1713, bsl2179, and bsl2063; Halobaculum marismortui rmAC1630 and rmAC385; and Mesorhizobium loti msl3517 and mrl3700) (see the file RRsseqUpdate in the supplemental material), have been submitted to the NCBI RefSeq database. The domain architecture of the Mlr3700 protein is discussed in the Results.

**RESULTS**

Response regulators encoded in prokaryotic genomes show a great variety of output domains and domain combinations. Unfortunately, owing to the relatively high sequence conservation of the REC domains, sequence database searches for relatives of certain response regulators often retrieve nonhomologous protein sequences that share only the REC domain (43). Hence, classification of response regulators has to be based primarily on their domain architectures and structures of the constituent domains. To achieve proper granularity, HTH domains of the response regulators have been grouped here by sequence similarity, following the approach utilized previously in the COG database (135). Response regulators that contained RNA-binding, protein-binding, or enzymatic output domains have been put into separate categories (Table 1). Any sequence fragments longer than 90 amino acid residues, not assigned to any known protein domain, were analyzed separately and, if appropriate, identified as members of existing or novel domain families. The results of this analysis are shown in Tables 1 and 2.

**Response regulators with DNA-binding output domains.** The great majority of all response regulators encoded in complete prokaryotic genomes combines the REC domain with a DNA binding HTH domain, typically either of NarL-type or of OmpR-type winged HTH family (Fig. 1). These domains are well studied and have known three-dimensional structures (12, 107). A factor of inversion stimulation (Fis)-type DNA binding domain (71, 73) is also commonly found in response regulators, either directly fused to the REC domain or containing an additional AAA+-type ATPase domain (Fig. 1). Response regulators of the latter type are referred to here as the NtrC family, after its best-studied representative (65, 76, 95, 145). The former type of domain architecture was first described in the RegA and PrrA response regulators from Rhodobacter sphaeroides (36, 73, 101, 115) and is referred to here as the PrrA family. Other widespread DNA-binding domains in re-
response regulators include the AraC-type HTH domain (106) and the LytTR domain (94), whose structure is still unknown (Table 1).

This work revealed a number of additional combinations of the REC domain with previously described DNA-binding domains, including cases where the HTH domain occupies the N-terminal part of the protein (Table 2). In some instances, newly identified output domains could be linked to known HTH domains by PSI-BLAST or CD searches. Thus, a previously uncharacterized N-terminal output domain in response regulators of the CV0537 family, encoded in an N-terminally uncharacterized domain in response regulators (Table 2).

TABLE 1. Widespread types of bacterial response regulators

<table>
<thead>
<tr>
<th>Response regulator</th>
<th>Output domain</th>
<th>Name</th>
<th>Pfam entry</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wHTH</td>
<td>PF00486</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTH</td>
<td>PF00196</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAA-FIS</td>
<td>PF00158 PF02954</td>
<td>76, 95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LytTR</td>
<td>PF04397</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIS</td>
<td>PF02954</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTH_AraC</td>
<td>PF00165</td>
<td>106</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANTAR</td>
<td>PF03861</td>
<td>119</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CheB</td>
<td>PF01339</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGDEF</td>
<td>PF00990</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAL</td>
<td>PF00563</td>
<td>46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-GYP</td>
<td>45, 111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP2C</td>
<td>PF07228</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HisKA-HATPase</td>
<td>PF00512 PF02518</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CheW</td>
<td>PF01584</td>
<td>110</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a The table lists response regulators with at least 40 representatives in the analyzed set, found in at least two bacterial phyla.

b Description of the domain structure, if available, or sequence alignment.

c This COG contains only the output domain of this response regulator.

d Includes 50 response regulators that contain both GGDEF and EAL domains.

Response regulators with enzymatic output domains. Recognition of the REC domain as a universal regulatory domain was prompted by sequencing of the chemotaxis protein CheB, which combines a REC domain with an N-terminal CsrA domain (50, 108), was found in Chromobacterium violaceum and Burkholderia spp. (beta-proteobacterial transcriptional regulator, BetR, deposited in the Pfam database as domain PF08667), was shown to be related to the XRE-type HTH domain (PF01381), found in Cro and cl repressors (see Fig. 1S in the supplemental material). In some cases, however, response regulators that were originally assumed to be DNA-binding showed no statistically significant similarity to any known DNA-binding domain. For example, nine haloarchaeal response regulators annotated as HxK1 through HxK6 (Table 2) contain a new output domain (renamed here HalX) (see Fig. S2 in the supplemental material, whose similarity, if any, to the homodomain-containing animal HxK proteins is totally spurious. All such response regulators have been classified as "uncharacterized." Most of these newly described domain combinations are relatively rare and have a narrow phylogenetic distribution (Table 2).

Response regulators with RNA-binding output domains. Of all the numerous RNA-binding domains described in the past several years (7), only one, ANTAR (119), has been commonly found in response regulators (Table 1). In transcriptional regulators such as the AmiR and NasT type, this domain stimulates transcription by preventing transcription termination at rho-independent terminators (96, 119). A single instance of a response regulator formed by another RNA-binding domain, a fusion of the REC domain with an N-terminal CsrA domain (50, 108), was found in Rhodopirellula baltica protein RB8820 (Table 2).
GYP domain has been proposed to function as a c-di-GMP hydrolase, based on its peculiar phyletic pattern, complementing that of the EAL domain (45, 46). Although this suggestion is consistent with participation of the HD-GYP-containing response regulators from mostly in gamma-proteobacteria. The PglZ domain, found in Dechloromonas aromatica, is often encoded on integrative plasmids (23). While its exact function is unknown, the PglZ protein and its close homologs often found at the C terminus of sensory histidine kinases, which are then referred to as “hybrid” kinases. The PglZ (phage growth limitation) gene of Streptomyces coelicolor A3 (2) is part of a four-gene operon that somehow controls phage propagation in this organism (128). Its homologs, which include Salmonella enterica serovar Typhimurium protein STM4492, are often encoded on integrative plasmids (23). While its exact function is unknown, the PglZ protein and its close homologs have all the hallmarks of the alkaline phosphatase protein superfamily (see Fig. S3 in the supplemental material). Given that members of this superfamily typically function as phosphatases, phosphomutases, or phosphodiesterases (44), PglZ is likely to have a similar enzymatic activity.

### Hybrid histidine kinases and REC-HisK response regulators

In addition to response regulators, the REC domain is often found at the C terminus of sensory histidine kinases, which are then referred to as “hybrid” kinases. The search for hybrid histidine kinases and REC-HisK response regulators in the Protein Data Bank (PDB) and other databases has led to the discovery of several new hybrid kinases. These kinases are characterized by the presence of a REC domain at the C terminus of the kinase, which is involved in dimerization and autophosphorylation. The REC domain is a highly conserved domain that is found in a variety of proteins, including response regulators, DNA repair enzymes, and metabolic regulators. The presence of the REC domain in hybrid histidine kinases suggests that they may have a role in regulating the activity of other proteins or signaling pathways.

### Table 2: Response regulators with unusual domain organizations

<table>
<thead>
<tr>
<th>Family name, domain organization</th>
<th>Total no.</th>
<th>Organism name</th>
<th>Gene name (accession no.)</th>
<th>Pfam entry</th>
<th>COG no.</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Bacillus subtilis</td>
<td>BSU24220 (CAB14353)</td>
<td>PB005154</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Caulobacter crescentus</td>
<td>CC3477 (AAK25439)</td>
<td>PF04542</td>
<td>PF04545</td>
<td>1595 26</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Vibrio cholerae</td>
<td>VCA0189 (AAF96102)</td>
<td>PF04509</td>
<td>1776</td>
<td>98 129</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Bradyrhizobium japonicum</td>
<td>Bbr2288 (BAC47553)</td>
<td>PF00211</td>
<td>2114</td>
<td>115</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Rhodobacter sphaeroides</td>
<td>Rso6853 (CAD14183)</td>
<td>PF00991</td>
<td>0455</td>
<td>77</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Nocardiococcus amyloliquefaciens</td>
<td>Nia2850 (BAD64127)</td>
<td>PF00070</td>
<td>2345</td>
<td>72</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Leifsonia xyli</td>
<td>Lexl99320 (AAT88820)</td>
<td>PF02585</td>
<td>2120</td>
<td>132</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Methanobacterium thermoautotrophicum</td>
<td>MTH548 (AA85054)</td>
<td>PF00355</td>
<td>1255</td>
<td>133</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Dechloromonas aromatica</td>
<td>Daro_0966 (AAZ25722)</td>
<td>PF00347</td>
<td>2804</td>
<td>102</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Parpyrophosphorobacter johnsonii</td>
<td>PG0928 (AAQ66063)</td>
<td>PF003865</td>
<td>This work</td>
<td></td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Vibrio cholerae</td>
<td>VCI0181 (AAF94240)</td>
<td>PF003868</td>
<td>1639</td>
<td>This work</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Vibrio cholerae</td>
<td>VCU0329 (CAD78826)</td>
<td>PB00069</td>
<td>0515</td>
<td>117</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Vibrio cholerae</td>
<td>VCU0329 (CAD78826)</td>
<td>PB00069</td>
<td>0515</td>
<td>117</td>
</tr>
<tr>
<td>DNA-binding, REC-HTH</td>
<td></td>
<td>Vibrio cholerae</td>
<td>VCU0329 (CAD78826)</td>
<td>PB00069</td>
<td>0515</td>
<td>117</td>
</tr>
</tbody>
</table>

**Example Output domain**

- **DNA-binding, REC-HTH**: Bacillus subtilis, Caulobacter crescentus, Vibrio cholerae
- **Response regulators**: Bacillus subtilis, Caulobacter crescentus, Vibrio cholerae
- **Phosphorylation enzymes**: Vibrio cholerae, Dechloromonas aromatica
- **Response regulators with unusual domain organizations**: Vibrio cholerae, Dechloromonas aromatica

**Reference(s)**

1. GALPERIN J. BACTERIOL. 4172
2. GALPERIN J. BACTERIOL. 4172
method used in this work retrieved hybrid histidine kinases as well as response regulators; these were subsequently separated based on their domain architectures. In *E. coli*, five histidine kinases (ArcB, BarA, EvgS, RcsC, and TorS) out of 30 are hybrid; in *B. subtilis*, there are no hybrid kinases with the exception of the YwpD protein that combines a truncated REC domain with a HisK domain (for simplicity, two separate subdomains of histidine kinases, the dimerization/phosphoacceptor one, HisKA, represented by Pfam entries PF00512, PF07568, PF07730, and PF02518, and the catalytic one, HATPase_c or PF02518, are treated here as a single HisK domain). However, the YwpD protein differs from typical hybrid kinases in that its REC domain is located at the N terminus, and there is no (other) sensory domain (Fig. 2). This means that the only signal that could be sensed by the YwpD protein is the phosphorylation (or other modification) of its N-terminal REC domain. Therefore, YwpD-like histidine kinases can be considered yet another form of response regulators with enzymatic output domains whose activity is modulated by appropriate sensory histidine kinases. Alternatively, the REC and HisK domains of such proteins might not be regulated at all and function solely as sinks for the phosphoryl residues. In any case, similarly to the response regulators of the PleD, VieA, and RpfG families, YwpD-like histidine kinases are members of complex signal transduction cascades operating inside the bacterial cell (41).

Unfortunately, the separation between YwpD-like response regulators and hybrid histidine kinases is not clear-cut. Genomic analysis reveals numerous intermediate forms with additional signaling domains, typically PAS and/or GAF, which could affect the enzymatic activity of the output domain (Fig. 2). For the purposes of this work, only YwpD-like histidine kinases without any discernible signaling domains intervening between REC and HisK were counted as bona fide response regulators.

The search for REC-containing proteins also revealed hybrid histidine kinases which, in addition to the REC domains, contain DNA-binding domains at their C-termini. These “one-component” signal transduction systems (141), exemplified by *B. cereus* protein BC3207, look like fusions of classical histidine kinases with full-length response regulators. Such proteins are encoded in a variety of bacterial genomes, sometimes in significant numbers. In *Bacteroides thetaiotaomicron*, for example, 36 out of 44 hybrid histidine kinases have tandem Arac-like DNA-binding domains at their C-termini (species-specific listings of hybrid histidine kinases are available in the files that are linked to Table S1 in the supplemental material). In *Bacteroides fragilis* strain YCH46, this domain organization is seen in 11 out of 17 hybrid histidine kinases. Some of such fusion proteins contain periplasmic, extracellular, or integral membrane sensory domains, so that they can only act as transcriptional regulators while their binding sites on the chromosomal DNA are positioned in the vicinity of the membrane.

**Response regulators with protein-binding domains.** In some response regulators, output domains are devoid of (known) enzymatic activity and exert their effects through protein-protein interactions. These include chemotaxis response regulators of the CheV type, combining REC with the protein-binding CheW-like domain (Table 1), and response regulators of the Hnr-type (see below), as well as response regulators that contain PAS, GAF, TPR, or HPt domains as the only output domains (Table 2). While PAS and GAF domains are known to contain periplasmic, extracellular, or integral membrane sensory domains, so that they can only act as transcriptional regulators while their binding sites on the chromosomal DNA are positioned in the vicinity of the membrane.

**Clade-specific response regulators.** While response regulators of OmpR, NarL, and many other families are widespread in *Bacteria*, some response regulators are found only in representatives of certain phylogenetic lineages. Some of them can be considered true synapomorphies—shared derived characteristics of the representatives of a particular group to the
exclusion of other groups—whereas others are encoded only in organisms with larger genomes and could have been lost in streamlined genomes of parasites and other small-genome organisms.

Spo0A. The master regulator of sporulation, Spo0A, is shared by all spore-forming representatives of Firmicutes, namely, the bacteria belonging to the orders Bacillales, Clostridiales, and Thermobacillales (122, 123). Remarkably, in the genome sequence of Clostridium perfringens strain 13, the spo0A gene is inactivated by what authors claim is a verified frameshift (118). If so, the sequenced strain should be unable to sporulate and produce toxins, which would make it a convenient laboratory strain but hardly a representative of the naturally occurring clostridia. In any case, Spo0A combines the REC domain with an unusual HTH domain that has additional \( /H_9251 \)-helices (78). Outside Firmicutes, this HTH domain has been seen so far only in the spore-forming high-GC gram-positive bacterium Symbio bacterium thermo-

<table>
<thead>
<tr>
<th>Domain Architecture</th>
<th>Source Organism</th>
<th>Gene Name</th>
<th>NCBI Protein Database Accession Number</th>
</tr>
</thead>
</table>
| FIG. 2. Domain architectures of response regulators with the histidine kinase (HisK) output domain. Domain symbols are drawn approximately to scale. Some of the domains are highly diverged and could be recognized only by searching the CDD database using relaxed stringency parameters. Source organisms, gene names, and NCBI protein database accession numbers for the domain architectures shown above are as follows (domain architecture number and accession number) B. subtilis ywpD (1; CAB05945) or Nostoc sp. PCC 7120 all3764 (1; BABA75463); Nostoc sp. PCC 7120 all1279 (2; BABA73236), all1968 (16; BABA73667), and all4097 (17; BABA75796); Bradyrhizobium japonicum bhi1030 (3; BAC46295); Halocella marismortui rma0374 (4; AAV45588), rma0379 (5; AAV48063), rma02533 (6; AAV47335), rma0075 (7; AAV45155), rma09848 (8; AAV45589), rma1626 (9; AAV46539), rma3361 (10; AAV48050), rma1495 (12; AAV46418), and pNG7156 (13; AAV44864); Synechocystis sp. PCC 6803 sbu0222 (11; BAA10222); Anabaena variabilis Ava_80391 (14; ABA24903); Gloeobacter violaceus gll0634 (15; BAC88575); Desulfurobacter vulgaris DUV2677 (18; AAS97149); Dechloromonas aromatica Daro_2847 (19; AA47577). Proteins with domain architecture numbers 1, 14, and 18 were considered bona fide response regulators.
phylum, which is currently assigned to the Actinobacteria but is apparently a member of the Firmicutes (140). Thus, Spo0A is a typical synapomorphy, common to the spore-forming bacteria of the Firmicute lineage.

YcbB. *B. subtilis* response regulator YcbB has been originally described as one of the genes whose disruption increased cell resistance to protonophores (105). This effect remained enigmatic until a recent study showed that YcbB controls the glutamine utilization operon (113). YcbB has been renamed GlnL, which is bound to cause confusion, as it is unrelated to the *E. coli* GlnL (NtrC) response regulator. In fact, the C-terminal domain of YcbB does not show obvious relation to any previously described HTH domain (see Fig. S4 in the supplemental material). Since DNA binding by YcbB has now been experimentally demonstrated (113), its C-terminal domain can be considered a new type of the DNA-binding HTH domain.

PatA. Response regulators of the PatA type, which control heterocyst development, phototaxis, and chromatic adaptation in cyanobacteria (18, 80), combine the C-terminal REC domain with N-terminal PATAN (PatA N terminus) domain and an HTH domain in the middle (83). Although the PATAN domain itself is found in a variety of bacteria, the PATAN-REC combination appears to be limited to cyanobacteria and γ-proteobacteria (83).

RpoE-REC. Rhizobia, brucellae, bartonellae, and other members of the alpha-subdivision of proteobacteria encode an unusual response regulator that combines a C-terminal REC domain with an N-terminal sigma-24 subunit of RNA-polymerase (σE or RpoE) (Table 2). With the sole exception of the marine bacterium *Silicibacter pomeroyi*, this response regulator is encoded in every alpha-proteobacterial genome that is larger than 1.6 Mb, often adjacent to the gene coding for a histidine kinase of the recently described HWE family (62). Sequence comparisons identified the RpoE-REC domain combination in the genome of *Mesorhizobium loti* (the *mlr3700* gene product), where the original genome annotation included only its C-terminal REC domain. Response regulators with the RpoE-REC domain combination have not been described in the available literature, and the only indication of their biological function(s) so far is the observation of increased expression of the *C. crescentus* response regulator CC3477 in the minimal medium compared to the rich medium (56). Given the DNA-binding properties of σE, these proteins can be predicted to serve as environmentally modulated transcriptional regulators. It remains to be seen whether their RpoE-like components could function as genuine sigma subunits of the RNA polymerase.

RssB. The Hnr-type response regulator, also referred to as RssB or SprE, regulates turnover of the stress sigma factor RpoS in *E. coli*, *Salmonella*, and several other enterobacteria (53). Due to the high degree of sequence conservation and narrow phyletic distribution, its output domain was previously considered to be unique (92). However, sequence analysis showed that this output domain is a degraded version of the PP2C-type Ser/Thr protein phosphatase that is shorter than the typical PP2C domain, has lost some of its Mn²⁺-binding Asp residues (Fig. 3), and is apparently devoid of the protein phosphatase activity. These observations are in line with the earlier suggestion (15) that RssB is a former anti-sigma factor that during evolution was recruited to serve as a recognition factor for proteolysis. However, RssB has apparently evolved from a protein phosphatase, not a protein kinase (an anti-sigma). As noted previously (69), phylogenetic distribution of this response regulator is limited to the gamma-proteobacterial families *Enterobacteriaceae* and *Vibrionaceae*, indicating that Hnr-dependent regulation of RpoS turnover is a peculiar feature of *E. coli* and its closest relatives. In pseudomonads, the degradation of the PP2C domain is much less pronounced, the key metal-binding residues are conserved, and the Hnr-related response regulators are likely to have the protein phosphatase activity (Fig. 3).

**REC-HDOD.** Several proteobacteria, including *Vibrio cholerae* and *Pseudomonas aeruginosa*, encode a response regulator with an unusual output domain, assigned to the COG1639 (41). In addition, this output domain is often encoded in the genomes of the same bacteria in a stand-alone form (see Fig. S5 in the supplemental material). The structure of one such protein, *Campylobacter jejuni* CJO248, has been solved by the Joint Center for structural Genomics and deposited in the Protein Data Bank under accession code 1VQR. Sequence and structural analysis of this output domain (deposited in the Pfam database as domain PF08668) showed its distant relationship to the widespread HD phosphohydrolase (PF01966) domain (9). This domain has been named the HD-related output domain, or HDOD, but the lack of conservation of the key metal-binding residues of the typical HD domain did not allow us to predict whether it still has a phosphatase or phosphodiesterase activity and, accordingly, what its role is, if any, in the signal transduction phosphorylation.

**Other response regulators.** Comparative analysis revealed several other clade-specific response regulators that are found exclusively in cyanobacteria (for example, the Slr2024 family), halobacteria (REC-HalX family), alpha-proteobacteria (CCO440 family), or epsilon-proteobacteria (CJ1608 family). A cyanobacteria-specific response regulator, Sll1879, combines the REC-domain with the Ycf55 domain, found in a stand-alone form in chloroplasts of green plants and red algae. No mention of experimental studies of these response regulators could be found in the available literature, and their functions remain unknown.

**Avoidance of certain response regulator types.** The availability of a complete genome sequence allows one to determine not only which proteins are encoded there but also which proteins (protein families) are missing. Absence of certain types of response regulators in all representatives of a given phylum could be an important feature of that phylum that might help in understanding mechanisms of signal transduction in its members.

Actinobacteria, for example, encode numerous transcriptional regulators of the SARP family (152), which combine an HTH domain with the BTAD domain, whose function is unknown (154). However, the current set of 390 actinobacterial response regulators does not include a single instance of a REC-SARP domain combination. The reasons for that remain unclear; other organisms, such as *Deinococcus radiodurans* and *Clostridium acetobutylicum*, do encode response regulators with this domain combination (Table 2). None of the available actinobacterial genomes encodes response regulators of the NtrC type, which is consistent with the absence of σ²⁴ in these
FIG. 3. Multiple sequence alignment of the Hnr-type output domain and its comparison with the protein phosphatase domains of the PP2C family. The top sequence is that of E. coli protein RssB (GenBank accession no. AAC74317). Conserved hydrophobic amino acid residues are shaded yellow; other conserved residues are shown in bold typeface. Metal-binding residues of PP2C-type protein phosphatases are shown in white.
domains of bacterial response regulators. Although actinobacteria encode signaling proteins with HisK, GGDEF, EAL, HD-GYP, and ACyc output domains, none of them is found in actinobacterial response regulators. Finally, only a small fraction of actinobacterial histidine kinases are hybrids. These features illuminate the peculiarity of the signal transduction in actinobacteria and suggest that representatives of this phylum utilize a rather straightforward mechanism of signal transduction that includes tight pairing of each HisK with a cognate DNA-binding (or RNA-binding) response regulator.

Archaeal transcriptional regulators and signal transducers are generally similar to those in bacteria (16, 47). However, archaeal response regulators differ dramatically from bacterial ones: no sequenced archaeal genome encodes OmpR, NarL, NtrC, LyrR, or PrrA family response regulators (see Table S1 in the supplemental material; see also reference 141). The only response regulators that are common for bacteria and archaea are the stand-alone REC domain and CheB-like and REC-HisK combinations. Other archaeal response regulators include REC-PAS and REC-GAF domain combinations, as well as combinations of the REC domain with uncharacterized lineage-specific domains, found, for example, only in halobacteria or only in metanomas.

Statistics of response regulators. The number of response regulators encoded in each complete microbial genome shows strong positive correlation with the genome size. In the logarithmic coordinates, this dependence shows a slope of approximately 1.8, meaning that it is close to the quadratic function (Fig. 4) (see also reference 141). There are certain deviations from the general trend: some representatives of beta-, delta-, and epsilon-proteobacteria, such as Geobacter sulfurreducens, Wolinella succinogenes, and Thiobacillus denitrificans, encode substantially more response regulators than other organisms with the same genome sizes. At the same time, other organisms encode relatively few response regulators or none at all (e.g., Sulfolobus spp.) (Fig. 4).

**DISCUSSION**

The sequence analysis of REC domain-containing response regulators presented in this work shows that functions of these proteins are not limited to transcriptional regulation and chemotaxis. While transcriptional regulators with DNA-binding output domains comprise about two-thirds of all response regulators, the remaining third consists of stand-alone REC domains, response regulators with enzymatic or protein-binding output domains, and response regulators whose output domains have not been characterized and whose functions remain unknown. In a somewhat similar analysis by Ulrich and colleagues (141), stand-alone REC domains and response regulators with unknown output domains were not considered, resulting in a much higher fraction of DNA-binding response regulators in their set.

Mechanisms of REC-mediated signaling. Stand-alone REC domains, such as chemotaxis regulator CheY and sporulation regulator Spo0F, have two principal modes of transmitting regulatory signals. Both depend on protein-protein interactions of the phosphorylated form REC–P with its various targets but differ in the fate of the phosphoryl group, which can stay on the REC domain or be transferred to another protein with REC–P functioning as an intermediate in a multicomponent phosphorelay (37). In some cases, REC–P might function just as a sink for phosphoryl groups (10, 130). In most bacteria, chemotactic signaling by CheY involves direct binding of its phosphorylated form, CheY–P, to the flagellar switch protein, FliM (24, 149). In archaea, which have an entirely different flagellar machinery that does not include FliM, CheY-like proteins interact with a different, still unidentified, protein (130). In cyanobacteria, whose motility apparatus is related to the type IV pili, CheY–P interacts with yet another type of switch complex (18). This is also the case for type IV pili-mediated social motility of Myxococcus xanthus (144). In addition to their role in chemotaxis, stand-alone REC domains participate in developmental processes, including regulation of sporulation in B. subtilis (Spo0F), heterocyst formation in Nostoc sp. (DevR), and cyst cell development in Rhodospirillum centenum (17, 27, 37). Furthermore, separately expressed REC and methylesterase domains of the CheB protein have been shown to form a complex even in the absence of a linker (6). These data suggest that REC domains could participate in a variety of protein-protein interactions with yet unknown target proteins.

Another big group of response regulators combines the REC domain with various enzymatic domains that are involved in signal transduction. Besides chemotactic methylesterase CheB and protein phosphatase CheC, these include the diguanylate cyclase (GGDEF), c-di-GMP phosphodiesterase (EAL and HD-GYP), adenylate cyclase, Ser/Thr-specific protein phosphatase, and histidine kinase output domains. In some cases, phosphorylation of the REC domain has been shown to activate the attached enzymatic domains (34, 99). While this activation could be due to the relief of enzyme inhibition by nonphosphorylated REC (34), it appears that in the case of CheB, both relief of inhibition and genuine activation take place (5, 6).

Another mechanism of REC-mediated signaling is the well-documented propensity of phosphorylated REC domains to form dimers. Phosphorylation-induced dimerization plays a key role in the regulation of DNA-binding response regulators, dramatically improving their binding to the palindromic or direct-repeat chromosomal binding sites. This mechanism has
been demonstrated for the response regulators of the OmpR, NarL, LytR, and PrrA families (12, 73, 88, 94). In the response regulators of the NtrC family, all three domains appear to participate in oligomerization, which regulates the activity of its AAA+ ATPase domain (67, 76).

The interplay of dimerization and relief-of-inhibition mechanisms could be particularly important for response regulators involved in c-di-GMP turnover. In response regulators of the REC-GGDEF type, phosphorylation of the REC domain activates the diguanylate cyclase by enabling formation of a GGDEF dimer, the active form of the enzyme (99, 112). In contrast, the c-di-GMP-specific phosphodiesterase (the EAL domain) is active as a monomer (114). Therefore, activation of EAL phosphodiesterase by REC phosphorylation in response regulators of the REC-EAL type (131, 137) appears to be due to the relief-of-inhibition mechanism. This distinction is particularly important for the response regulators of REC-GGDEF-EAL domain architecture, which comprise 1.1% of all response regulators in the analyzed set. Although in many such proteins either the GGDEF or EAL domain appears to be inactivated (114), there are some without obvious deviations from the consensus (109). Such response regulators should have the only the phosphodiesterase activity in their monomeric form but might acquire diguanylate cyclase activity upon dimerization. It is entirely possible that upon phosphorylation of their REC domains, the overall activity of such response regulators changes from c-di-GMP-hydrolase to c-di-GMP synthetase and vice versa.

In addition to dimerization, relief-of-inhibition, direct activation, and protein-protein interactions with or without phosphoryl transfer, REC-mediated signaling might involve other mechanisms, such as, for example, allosteric regulation of the output domains. This versatility of the REC domain is consistent with the diversity of signal output domains that interact with it.

New types of response regulators. The same regulatory mechanisms mentioned above—dimerization, protein-protein interaction, and activation/relief-of-inhibition—can be expected to work in the newly described response regulators. Among the new types of DNA-binding response regulators (Table 2), DNA-binding domains of LysR, MerR, and HxI regulators are known to function as dimers (25, 93, 155), as is the RNA-binding CsrA domain (50). DNA-binding domains of the XRE-family, which includes cl and Cro regulators, also function as dimers, suggesting that the BetR domain, a newly described distant member of this family, would function the same way. However, the RpoE7-like domain of the
CC3477-type (RpoE-REC) response regulators is not known to form dimers, so they most likely act through relief of inhibition and/or protein-protein interactions of their C-terminal REC domains.

Among enzymatic output domains of response regulators (Tables 1 and 2), adenylate and diguanylate cyclases, histidine kinases, and RsbW-type Ser/Thr protein kinases (anti-sigma factors) are known to function as dimers (19, 28, 82, 156). In these response regulators, both phosphorylation-induced dimerization of the REC domains and relief-of-inhibition could contribute to the activity of the output domains. In all other cases, a nonphosphorylated REC domain can be expected to block enzyme activity the same way it does that in the other cases, a nonphosphorylated REC domain can be expected to block enzyme activity the same way it does that in the CheB structure (34). However, just as in CheB, a direct enzyme activation by REC~P, as well as an allosteric regulation, also remains a possibility.

Phylogenetic distribution of response regulators. The census of response regulators confirmed several trends seen previously in the census of signal transduction proteins (42). In accordance with previous observations (46, 68, 141, 142), the total number of response regulators encoded in completely sequenced prokaryotic genomes positively correlates with the genome size (Fig. 4) and the total number of encoded proteins. The largest deviation from the general trend is seen in the same representatives of beta-, delta-, and epsilon-proteobacteria that have been previously found to encode more signaling proteins than others with the same genome size (42). Given only a minimal overlap between the two data sets, these observations show that certain environmental bacteria encode a disproportionately higher number of sensor, signal transduction, and response regulator proteins than other prokaryotes with the same genome size. On the other end of the spectrum are bacteria and archaea that encode fewer response regulators than the average bacterium of the same genome size. These are typically obligate parasites, such as *Yersinia pestis*, or organisms such as *Sulfolobus spp.*, or *Thermoplasma spp.* that inhabit relatively stable and unique environments where few parameters, if any, ever change. These data validate the use of the term “bacterial IQ” (42) to describe the fraction of the bacterial proteome dedicated to environmental sensing and signal transduction.

Origin and function of diverse domain architectures. The diversity of output domains in response regulators (Tables 1 and 2) suggests that virtually any DNA-binding or enzymatic domain could be fused to the REC domain to form a response regulator. For DNA-binding domains, this is probably correct, as representatives of all major classes of the HTH domain (8) could be found in response regulators (Table 2). As discussed above, the propensity of the REC domain to dimerize upon phosphorylation offers an effective mechanism of transcriptional regulation. Remarkably, HTH domains such as LysR or MerR that are typically localized at the N-termini of the corresponding transcriptional regulators have the same position in response regulators (Table 2), a clear deviation from the standard REC-HTH architecture. This diversity makes it even more difficult to explain the predominance of OmpR-type and NarL-type response regulators, far fewer numbers of the LysR-type, PrrA-type, and YesN-type response regulators, and only sporadic appearance of the response regulators with other types of the HTH domain (Fig. 1 and Tables 1 and 2).

Most enzymatic output domains found in response regulators (Tables 1 and 2) are signal transduction domains themselves, often found in transmembrane receptors (42). Such fusions add yet another layer of regulation in the complex intracellular signal transduction network, allowing modulation of the cellular levels of cAMP and c-di-GMP, as well as the activity of Ser/Thr protein kinases and phosphatases, by sensory histidine kinases in response to specific environmental cues (41). Fusions with other enzymes typically have a narrow phylogenetic distribution, and it is not known if they are ever expressed. Such combinations must have arisen from random gene translocation and fusion events and were later fixed in evolution because they allowed tight regulation of the enzyme activity. In one such case, a fusion of the REC and thiorodoxin reductase domains could play an important role in environmental regulation of the cellular dithiol-disulfide ratio. This domain combination has been found in several actinobacteria and cyanobacteria and in *Solibacter usitatus*, a recently sequenced representative of the phylum *Acidobacteria*. Plant pathogen *Leifsonia xyli* encodes a unique response regulator with the LmbE output domain, recently shown to have N-acylglucosamine deacetylase activity (Table 2). Such domain combination could allow this bacterium to regulate degradation of N-acylglucosamine-containing plant polymers, such as lignin and chitosan, depending on the environmental conditions.

In general, it appears that fusing an existing metabolic enzyme to the REC domain is an evolutionarily harmless procedure: the resulting fused protein is inactive unless its REC domain is phosphorylated by a cognate histidine kinase. The same logic could be applied to the fusions between REC domains and other signaling, ligand-binding, or protein-binding domains. Indeed, the large number of rare domain fusions (Table 2; see Table S1 in the supplemental material) supports the view that signaling domain fusions follow the Lego principle: almost any fusion can occur but only some of them remain fixed in the evolution (43). If this logic is correct, many more types of response regulators can be expected to be found in microbial genomes. Several relatively common, previously uncharacterized rare or unique domains identified in this study are listed in the supplemental material and will soon be available in the public domain databases. An experimental study of these novel response regulators and analysis of their potential functions will remain the task for future work, which will provide further understanding of the organization and evolution of the prokaryotic signal transduction systems.

ACKNOWLEDGMENTS

I thank Andrei Gabrielian, Bill Klimke, Darren Natale, Tao Tao, and Yuri Wolf for helpful advice; Mark Gomelsky, Igor Zhulin, and two reviewers for critical comments; and many other colleagues for suggestions.

This study was supported by the Intramural Research Program of the National Library of Medicine at the U.S. National Institutes of Health.

REFERENCES


123. Volz, K. 1993. Structural conservation in the CheY superfamily. Biochem-


126. Welch, M., K. Oosawa, S. Aizawa, and M. Eisenbach. 1993. Phosphoryla-


129. Whitcomb, R. L., and M. Y. Galperin. 1996. Con-