Published in final edited form as:
Methods Enzymol. 2007 ; 422: 1–31. doi:10.1016/S0076-6879(06)22001-9.

Comparative Genomic and Protein Sequence Analyses of a Complex System Controlling Bacterial Chemotaxis

Kristin Wuichet, **Roger P. Alexander**, and **Igor B. Zhulin**

Abstract

Molecular machinery governing bacterial chemotaxis consists of the CheA-CheY two-component system, an array of specialized chemoreceptors, and several auxiliary proteins. It has been studied extensively in *Escherichia coli* and, to a significantly lesser extent, in several other microbial species. Emerging evidence suggests that homologous signal transduction pathways regulate not only chemotaxis, but several other cellular functions in various bacterial species. The availability of genome sequence data for hundreds of organisms enables productive study of this system using comparative genomics and protein sequence analysis. This chapter describes advances in genomics of the chemotaxis signal transduction system, provides information on relevant bioinformatics tools and resources, and outlines approaches toward developing computational framework for predicting important biological functions from raw genomic data based on available experimental evidence.

Introduction

Signal transduction systems link internal and external cues to appropriate cellular responses in all organisms. Prokaryotic signal transduction can be classified into three main families based on the domain organization and complexity: one-component systems, classical two-component systems anchored by class I histidine kinases, and multicomponent systems anchored by class II histidine kinases often referred to as chemotaxis systems (Bilwes *et al*., 1999; Dutta *et al*., 1999; Stock *et al*., 2000; Ulrich *et al*., 2005). As their name suggests, one-component systems consist of a single protein that is capable of both sensing a signal and directly affecting a cellular response, either through a single domain (such as a DNA-binding domain that senses a signal through its metal cofactor) or multiple domains (separate input and output domains) (Ulrich *et al*., 2005). As a consequence of their single protein nature and typical lack of transmembrane regions, one-component systems are predicted to primarily sense the internal cellular environment, while the division of input and output between two or more proteins and association of the sensor with the membrane in two-component systems allows them to detect both internal and external signals (Ulrich *et al*., 2005). The chemotaxis system centered around the class II histidine kinase CheA contains multiple proteins separating input and output, along with additional regulatory components that are not present in class I histidine kinase containing two-component systems. There are many common input (sensing) modules among all three families of prokaryotic signal transduction; one-component systems and two-component systems also share common outputs (Ulrich *et al*., 2005), whereas two-component systems and chemotaxis systems share several common signaling modules (Dutta *et al*., 1999; Stock *et al*., 2000).

The chemotaxis system is classically portrayed as a network of interacting proteins, which senses environmental stimuli to regulate motility. The system consists of two distinct pathways: an excitation pathway that has the downstream result of interacting with the motility organelle and an adaptation pathway that provides a mechanism for molecular memory (Baker *et al*., 2006; Wadhams and Armitage, 2004). The excitation pathway involves methyl-accepting chemotaxis proteins (MCPs) for sensing environmental signals that are transmitted to a scaffolding protein, CheW, and a histidine kinase, CheA, via a highly conserved cytoplasmic signaling module of the MCPs. The signals regulate the kinase activity of CheA and the phosphorylation state of its cognate response regulator CheY controls its affinity for the motor. Many chemotaxis systems have one or more phosphatases (CheC, CheX, and/or CheZ) involved in the excitation pathway that aid in dephosphorylating CheY (Szurmant and Ordal, 2004). Signal propagation through the MCPs is further controlled in most systems by an adaptation pathway that regulates their methylation state via the CheB methylesterase, a response regulator that is phosphorylated by CheA to stimulate the removal of methyl groups from the receptors, and the CheR methyltransferase that constitutively methylates specific glutamate residues of the receptors. Many chemotaxis systems have an additional adaptation protein, CheD, for the deamidation of particular amino acid side chains of many MCPs prior to their methylation, and in some of these systems CheD also interacts with CheC to increase its dephosphorylation activity (Chao *et al.*, 2006; Kristich and Ordal, 2002). The final characterized chemotaxis protein is CheV, a fusion of CheW and a CheY-like receiver domain, which affects the signaling state of the MCP based on its phosphorylation state as controlled by the CheA kinase (Karatan *et al.*, 2001; Pittman *et al.*, 2001).

In addition to component diversity between chemotaxis systems, there are also functional differences between their outputs. Historically, the focus of detailed molecular investigation is on the chemotaxis system that controls flagellar motility, but studies have demonstrated that chemotaxis systems are also involved in regulating type IV pili-based motility (Bhaya *et al.*, 2001; Sun *et al.*, 2000; Whitchurch *et al.*, 2004). Even more recently, chemotaxis systems were implicated in controlling diverse cellular functions, such as intracellular levels of cyclic di-GMP, transcription, and other (Berleman and Bauer, 2005; D'Argenio *et al.*, 2002; Hickman *et al.*, 2005; Kirby and Zusman, 2003). Many organisms have multiple chemotaxis systems that can have both overlapping and/or unrelated functional outputs (Berleman and Bauer, 2005; Guvener *et al.*, 2006; Kirby and Zusman, 2003; Martin *et al.*, 2001; Wuichet and Zhulin, 2003). Beyond the functional diversity of the system outputs, there can be significant mechanistic diversity within these functional classes. For example, the signaling and adaptation mechanisms in *Escherichia coli* and *Bacillus subtilis* differ markedly. In *E. coli*, positive stimuli inhibit CheA activity, whereas in *B. subtilis* the opposite is true. In *E. coli*, MCP demethylation increases in response to negative stimuli only, whereas in *B. subtilis*, it occurs in response to both positive and negative stimuli (Szurmant and Ordal, 2004).

The diversity found among chemotaxis systems cannot be efficiently addressed by experimental means alone, nor can the questions about the function and origin of this system. Initial genomic studies have already identified the core set of chemotaxis proteins as CheA, CheW, CheY, and MCP, which are present in all chemotaxis systems (Zhulin, 2001), unlike the sporadic distributions of CheC, CheD, and CheZ (Kirby *et al.*, 2001; Szurmant and Ordal, 2004; Terry *et al.*, 2006) and the occasional absence of CheB and CheR (Terry *et al.*, 2006; Zhulin, 2001). Diversity within the CheA domain organization was also reported (Acuna *et al.*, 1995; Bhaya *et al.*, 2001; Whitchurch *et al.*, 2004), as well as the broad repertoire of MCP sensor domains (Aravind and Ponting, 1997; Shu *et al.*, 2003; Taylor and Zhulin, 1999; Ulrich and Zhulin, 2005; Zhulin, 2001; Zhulin *et al.*, 2003) and their evolutionary trends (Wuichet and Zhulin, 2003), and the length variability of the MCP signaling module (LeMoual and Koshland, 1996). Motivating factors to further study the chemotaxis system using comparative genomic methods are the wealth of genomic data available for prokaryotes, the large evolutionary distances between prokaryotes that have this system, and the propensity for its components to be encoded in gene clusters. The extensive molecular and biochemical characterizations of the system and its components and the availability of three-dimensional structures for most of the components provide most valuable information for comparison and validation of findings obtained through computational analysis. Although this chapter focuses on the chemotaxis system, the methodology of this research is applicable to all signal transduction systems, prokaryotic and eukaryotic, with the caveats that certain thresholds (e.g., sequence conservation) must be altered to suit the evolutionary rate of a given protein or domain and that some techniques (e.g., gene neighborhood analysis) are best applied to prokaryotic systems.

Bioinformatics Tools and Resources for Identifying and Analyzing Chemotaxis Components

Many tools and databases are available to aid comparative genomic analyses. The SMART (Letunic *et al.*, 2004) and Pfam (Finn *et al.*, 2006) databases are primary sources for Hidden Markov Models (HMMs) that can identify conserved domains and domain combinations within protein sequences. Each model captures the key sequence features of a specific domain, based on the multiple alignments from which it is built. When a model for a given domain is not available or is inadequate (e.g., poor quality, artificial relationship between sequences), the sequence of a representative protein family member can be used to search against common sequence databases such as those at the National Center for Biotechnology Information (NCBI) (Wheeler *et al.*, 2006) using various versions of the Basic Local Alignment Search Tool (BLAST) algorithm (McGinnis and Madden, 2004). For comparative analysis we often focus on completely sequenced genomes, which make the RefSeq and microbial databases of NCBI (Wheeler *et al.*, 2006) ideal to search against, but even within these searches there are many ways to further narrow down search results, for example, by retrieving only sequences of a certain length range (McGinnis and Madden, 2004). While a single search iteration is standard in BLAST, the Position-Specific Iterative BLAST (PSI-BLAST) program enables iterative searches by updating a position-specific score matrix (PSSM) with each iteration. PSI-BLAST enables identifying many divergent members of a particular protein family (Altschul *et al.*, 1997). Typical sequence similarity searches compare DNA to DNA, DNA to protein, protein to DNA, or protein to protein. Ideally protein-to-protein searches should be performed because the greater number and diversity of sequence characters in proteins (20 amino acids versus 4 nucleotide bases) make them more sensitive. Searching a database using a PSSM output by PSI-BLAST produces more sensitive results than a search using a single protein sequence. This approach can be useful for searching protein family members in newly released genomes.

Although BLAST and PSI-BLAST searches are invaluable in comparative genomic analyses, they can be time-consuming and the results need to be analyzed carefully (which is particularly true for PSI-BLAST analyses). HMM domain models of well-characterized protein families can quickly identify new family members and help understand the functions of all proteins in the family. Regularly updated databases of protein domain architectures, such as SMART and Pfam, are important tools to begin to understand protein function at the individual or family level. The recently developed Microbial Signal Transduction (MiST) database expands this concept by extracting signal transduction profiles for all complete microbial proteomes, taking advantage of both SMART and Pfam models and the wealth of knowledge generated in the area of microbial signal transduction (Ulrich and Zhulin, 2007; Ulrich *et al.*, 2005). Figure 1 shows representative members of each chemotaxis protein as visualized in MiST, and the following sections discuss the best way to identify each of these proteins in public databases and other sources of genomic data.

Once protein family members are identified, other bioinformatics tools need to be employed in order to derive meaningful information about their function and relationships. Because multiple sequence alignments are the essential backbone of most comparative analyses, building high-quality multiple alignments is critically important. There are many programs currently available to build initial multiple alignments, including most popular Clustal (Chenna *et al.*, 2003), MUSCLE (Edgar, 2004), PCMA (Pei *et al.*, 2003), and T-COFFEE (Notredame *et al.*, 2000). ClustalW, MUSCLE, and PCMA are very fast programs suitable for a large number of sequences, particularly for a set of sequences that is highly conserved. T-COFFEE is slower, but has a higher accuracy for the alignment of sets of sequences that are not highly conserved. For a given set of sequences one program may produce better initial results than the others, but manual analysis and editing are then needed in most cases. Manual editing with a program such as SeaView (Galtier *et al.*, 1996) can help resolve the gap regions and poorly conserved alignment regions that are not handled well by the alignment software. The VISSA program aids manual editing by visualizing the secondary structure of each protein sequence in a given multiple alignment (Ulrich and Zhulin, 2005). Although some regions of a protein may display poor sequence conservation, there is often still pressure on these regions to maintain their secondary structure. Identifying unstructured regions that link secondary structure elements can also aid in the placement of gaps during the editing process.

A multiple sequence alignment provides immediate information about potentially important functional regions and individual amino acids by revealing highly conserved positions. The CONSENSUS script (http://coot.embl.de/Alignment/consensus.html) and the WebLogo program (Crooks *et al.*, 2004) can further aid in the identification of highly conserved positions in multiple alignments. If the structure of a protein is available, conserved sites can be easily visualized on the structure with structure viewing packages, such as DeepView (Schwede *et al.*, 2003) and PyMol (http://www.pymol.org). In order to cluster related protein sequences, phylogenetic trees can be built from a multiple alignment with many different methods and programs. The MEGA program (Kumar *et al.*, 2004) is a user-friendly tool used to build neighbor-joining trees for the quick identification of protein subfamilies based on sequence similarity. MEGA can also be used to easily view and edit trees produced by different multiple alignment programs. Sequence similarity is not always a reflection of the evolutionary history of a protein, as there can be multiple mutation events that obscure origins. For more precise evolutionary analysis, maximum likelihood trees are more appropriate and can be built using the ProML program of the PHYLIP package (Felsenstein, 1989). Because trees can often vary depending on the methods used to build them, it is best to validate them by independent means. Unless the gene encoding a protein is a subject of a frequent horizontal transfer, it is expected that most closely related proteins of a tree will be from closely related organisms. We also expect proteins with similar domain architectures to cluster together in a tree. Most importantly, because chemotaxis genes are encoded in conserved gene clusters, closely related proteins are predicted to be encoded in similar gene neighborhoods. Gene neighborhood (or genome context) analysis (Overbeek *et al.*, 1999) can become a very useful approach in elucidating specific interactions when multiple chemotaxis systems are encoded within a genome. Occasionally, distinct protein subfamilies can be correlated to specific motifs within the alignment as well as insertions or deletions that might be specific to their structure and function.

Defining MCP Membrane Topology

Methyl-accepting chemotaxis proteins are the receptors at the beginning of the chemotaxis signal transduction cascade that process environmental and intracellular sensory (input) signals and alter the activity of the CheA histidine kinase. MCP sequences typically consist of a sensory domain, a HAMP linker domain, and a signaling domain that interacts with CheA (Fig. 1). The HAMP and signaling domains are always cytoplasmic, but the membrane topology of the sensory domain varies. Figure 2 shows the four main classes of MCP membrane topology (Zhulin, 2001). Sensory class I MCPs have a periplasmic sensory domain anchored by an N-terminal transmembrane (TM) helix and connected by an internal TM helix to the HAMP linker and signaling domains. Most MCPs, including the Tar, Tsr, Trg, and Tap receptors of *E. coli*, have this sensory topology (Ulrich and Zhulin, 2005;Zhulin, 2001). Sensory class II MCPs have an N-terminal cytoplasmic sensory domain connected by an internal TM helix to the HAMP linker and signaling domains. The Aer aerotaxis receptor of *E. coli* is an example of a class II sensor (Bibikov *et al.*, 1997;Rebbapragada *et al.*, 1997). Since the previous classification of MCP sensor classes (Zhulin, 2001), many more MCP sequences have become available, and we now split sensor class III into two subgroups. Sensory class IIIc MCPs are anchored at their N terminus by a TM helix, downstream of which are a cytoplasmic sensory domain and the HAMP linker domain and cytoplasmic signaling domain. Sensory class IIIm MCPs are like class IIIc MCPs except that the sensory domain is membrane bound rather than cytoplasmic. The Htr8 aerotaxis receptor of *Halobacterium salinarum* (Brooun *et al.*, 1998) is an example of a sensory class IIIm receptor. Some MCPs are hybrids of class II and class III, containing a periplasmic sensory domain separated by a TM helix from an additional cytoplasmic sensory domain (Wuichet and Zhulin, 2003). Sensory class IV MCPs are entirely cytoplasmic; they lack TM helices and usually also HAMP domains. The oxygen sensor HemAT from *B. subtilis* is an example of a class IV sensor (Hou *et al.*, 2000;Zhang and Phillips, 2003).

Methyl-accepting chemotaxis protein sensor class and membrane topology can be easily determined by visual inspection of a two-dimensional domain model that includes TM regions (Fig. 2B). Transmembrane regions can be identified in MCPs and other proteins using various TM prediction programs. In our analyses, we mostly use Phobius (Kail *et al.*, 2004) and DAS-TMfilter (Cserzo *et al.*, 2002); they give similar results and are amenable to high-throughput scripting. It should be noted that because DAS-TMfilter is a modification of the Dense Alignment Surface (DAS) algorithm (Cserzo *et al.*, 1997) to screen out false positives, if one suspects “underpredicting” TM regions in an MCP of interest, the original DAS algorithm can be used on a case-by-case basis. Both DAS and Phobius can generate graphical TM prediction plots for visual inspection.

Diversity of Input (Sensory) Domains in MCPs

The MCP signaling domain is highly conserved because it maintains multiple protein-protein interactions within the chemoreceptor-kinase complex. MCP sensory domains, however, evolve rapidly, being a subject of frequent domain birth and death events, and are quite variable in sequence (Wuichet and Zhulin, 2003). In fact, the lack of good sensory domain models is still a unsolved problem not only in chemotaxis, but in microbial signal transduction in general (Ulrich and Zhulin, 2005). Figure 3 shows an array of well-defined sensory domains found in MCPs. PAS (Taylor and Zhulin, 1999) and GAF (Aravind and Ponting, 1997) are ubiquitous sensory domains of the similar protein fold (currently known simply as the PAS/GAF fold) that can be found throughout the prokaryotic and eukaryotic signal transduction. Most members of these domain familiesare cytoplasmic, although a divergent PAS subfamily is exclusively extracellular (Reinelt *et al.*, 2003). In addition to MCPs where they are located exclusively extracellularly, Cache family domains are also found in extracellular subunits of eukaryotic Ca^{2+} channels that are implicated in signal transduction (Anantharaman and Aravind, 2000). For some sensory domains, their signal specificity can be proposed in a narrow range, for example, the nitrate- or nitrite-responsive NIT domain (Shu *et al.*, 2003); however, in most instances, the MCP signal spectrum cannot be readily predicted by the sequence conservation of their sensory domains. Furthermore, for a significant number of MCPs, while the sensory topology can be determined from the pattern of TM helices, no known domains are identified by current domain models. These regions contain either known domains that are not recognized by low-sensitivity models or novel, uncharacterized domains. Further computational and experimental work is necessary to identify and understand the function of novel sensory domains in MCPs.

HAMP Domain Identification

The HAMP linker domain is an important module, which is present in many membrane-bound signal transduction proteins, including MCPs and the sensor histidine kinases of two-component systems (Aravind and Ponting, 1999). The HAMP domain is about 60 amino acids long and consists of two amphipathic *α* helices (AS1 and AS2) separated by a loop. Because of its structural flexibility, the mechanism of signal transmission by the HAMP domain has been difficult to characterize (Williams and Stewart, 1999); however, the nuclear magnetic resonance structure of a stable archaeal HAMP domain has been determined (Hulko *et al.*, 2006) and should lead to new developments in the field. Because some MCPs contain multiple HAMP domains, understanding of its mechanism should involve modularity and the possibility of self-interaction.

The domain models of the HAMP linker domain in Pfam and SMART (Pfam, HAMP; SMART, HAMP) are slightly different from each other and are of relatively poor quality. The Pfam domain model includes a fully lipophilic helix at its N terminus upstream of AS1, which overlaps the TM regions that determine MCP sensory topology (see example shown in Fig. 1). The SMART HAMP domain model extends three residues past the Pfam model, which is important to keep in mind when trying to establish the boundaries of the signaling domain. Most importantly, both the Pfam and the SMART domain models fail to identify HAMP domains in many sequences, where they are obviously present (Figure 4). If the domain organization of an MCP of interest contains a short region free of identified domains downstream of the membrane and upstream of the signaling domain, a PSI-BLAST search should be performed, which in many instances will lead to the detection of the HAMP domain.

MCP Signaling Domain

The cytoplasmic signaling domain of MCPs is a coiled coil with a hairpin at its base that is highly conserved in sequence. The presence of this highly conserved domain (HCD) in MCPs makes it possible to extract all MCP sequences from a genome with high confidence using the Pfam or SMART domain models of the cytoplasmic signaling domain (Pfam, MCPsignal; SMART, MA). It is important to bear in mind that the Pfam and SMART domain models do a poor job of delineating the exact boundaries of the signaling domain because of the significant variability of its length. LeMoual and Koshland (1996) identified three classes of a MCP signaling domain that were different in length by multiples of seven residues, or exactly two turns of an *α* helix in a coiled coil protein. The signaling domains of MCPs from *E. coli* have four 14-residue gaps relative to those from *B. subtilis*, a total of 56 residues difference in length. Most recent in-depth computational analysis resulted in the identification of seven major and several minor length classes of the MCP signaling domain revealing the subdomain organization and unusual evolutionary history of this important signaling module (Alexander and Zhulin, 2007).

MCP Pentapeptide Tether

Alexander and Zhulin (2007) collected 2125 MCP sequences from 152 bacterial and archaeal genomes and analyzed their C-terminal five residues. In *E. coli*, this C-terminal pentapeptide has been shown to bind to the adaptation enzymes CheB and CheR. The pentapeptide motif in *E. coli* MCPs is NWETF, but it was found that the motif could be generalized with an emphasis on two aromatic residues (-x-[HFWY]-x(2)-[HFWY]-). Only 217 MCPs from 67 of 152 genomes contained sequences that matched this motif. All of these MCPs belong to the same major class of the signaling domain as the five MCPs of *E. coli*. All but two of the organisms where pentapeptide-containing MCPs are found are proteobacterial, implying that the pentapeptide tether is a recently evolved mode of interaction between MCPs and adaptation enzymes. The pentapeptide can be easily identified in newly available MCP sequences by visual inspection or simple scripting in Perl.

The CheA Histidine Kinase: Domain Organization, Conservation, and Diversity

The CheA histidine kinase is an essential component of the chemotaxis system and has a complex multidomain architecture (Bilwes *et al.*, 1999; Stock *et al.*, 2000). Five domains were identified in CheA from model organisms *E. coli* and *B. subtilis*, but analysis of CheA sequences from more recent experimental studies have revealed that its domain architecture can be highly variable (Fig. 5) (Acuna *et al.*, 1995; Bhaya *et al.*, 2001; Porter and Armitage, 2004; Whitchurch *et al.*, 2004). CheA has a conserved core of four domains, a histidine phosphotransfer domain (Pfam, Hpt; SMART, HPT) that is autophosphorylated by ATP (Kato *et al.*, 1997), a dimerization domain (Pfam, H-kinase_dim) (Bilwes *et al.*, 1999), an ATPase domain (Pfam and SMART, HATPase_c) (Bilwes *et al.*, 1999), and a CheW scaffolding domain (Pfam and SMART, CheW) that is homologous to the CheW protein (Bilwes *et al.*, 1999). Although the dimerization domain (Pfam, H-kinase_dim) was not initially identified in some of the CheA proteins, this is the result of a poor domain model, as revealed by a multiple alignment of CheA sequences. The crystal structure of the dimerization, ATPase, and CheW domains of the CheA from *Thermotoga maritima* (Bilwes *et al.*, 1999) shows that the domain model does not cover the full domain length. Despite the dimerization domain discrepancy, the two-dimensional domain model for CheA does capture the overall information about the three-dimensional state of the protein (Fig. 6). Domain searches for proteins with both the HATPase_c and the CheW domains should easily identify CheA homologs without the need for additional search tools.

While the majority of CheA proteins contain all of the core domains, a few chemotaxis systems have the HPT domain detached from the other three core domains as a separate protein. An unusual split CheA in *Rhodobacter sphaeroides* has both parts found in the same gene neighborhood (Porter and Armitage, 2004). One gene encodes the dimerization domain, ATPase, and CheW domains while the other has the HPT and CheW domains. Neither protein is able to undergo autophosphorylation, but they are able to interact together for transphosphorylation. In *Synechocystis* sp. PCC 6803, the HPT domain protein is found in a region of the genome distant from its partner protein, which contains dimerization, ATPase, CheW, and REC (response regulator receiver) domains (Yoshihara *et al.*, 2002). Both proteins are necessary for chemotaxis similarly to the split CheA in *R. spharoides*. Although the core domains of CheA are conserved in all chemotaxis systems, the extensive domain architecture diversity implies that there is a high level of functional and mechanistic differences that must be addressed by comparative analysis and experiment.

The P2 domain (Pfam, P2) is absent from many CheA proteins, while many CheA have an additional carboxyl-terminal REC domain (Acuna *et al.*, 1995; Bhaya *et al.*, 2001; Whitchurch *et al.*, 2004). Histidine kinases that contain the REC domain are termed “hybrid.” For any CheA sequences that have extended undefined regions, PSI-BLAST searches should be performed in order to find out whether they contain known domains missed by current models or novel domains. Such searches identified multiple divergent HPT domains, and new domains not previously associated with CheA proteins, in undefined regions of a CheA homolog in *Pseudomonas aeruginosa* that regulates the type IV pili-based motility (Whitchurch *et al.*, 2004). Because the P2 domain has particularly low sequence conservation in comparison to the rest of the CheA domains, undefined regions between HPT and dimerization domains that are not characterized by low complexity are potential P2 domains. Crystal structures of the P2 domains of *T. maritima* and *E. coli* CheA proteins show that the *E. coli* P2 domain is reduced in size and missing some structural elements present in the P2 domain of *T. maritima* (McEvoy *et al.*, 1998; Park *et al.*, 2004a). Our sequence analysis of P2 domains and unidentified regions between HPT and dimerization domains revealed three classes of the P2 domain. Class I is represented by the *T. maritima* P2 domain. Class II shows structural similarities to class I, but it has an extra insertion. Class II also shows sequence similarities to the reduced P2 domain of class III, which is represented by the *E. coli* P2 domain. The three classes can be aligned accurately with the help of VISSA (Ulrich and Zhulin, 2005), and the gap regions of the alignment clearly show the three classes (Fig. 7). Despite the poor domain model for P2, all CheA proteins that contain a P2 domain have the typical domain architecture shown in Fig. 1 with the exception of the unusual HPT-CheW protein of *R. sphaeroides* (Porter *et al.*, 2002), where we find a previously unidentified P2 domain in the long undefined region between the two domains, and some archaeal CheA proteins that have tandem P2 domains.

Interestingly, many CheA proteins that lack P2 domains contain the C-terminal REC domain (Fig. 5). Conversely, none of the CheA proteins that contain the C-terminal REC domain has the P2 domain (the P2 domain was proposed in the hybrid CheA from *Synechocystis* [Bhaya *et al.*, 2001]; however, our computational analysis failed to support it). This observation can lead to several experimentally testable hypotheses.

The CheY Response Regulator: Big Problems of the Small Protein

Although essentially all CheY proteins can be identified by domain searches, such searches cannot identify CheY proteins exclusively because there is no specific domain model for CheY. CheY is a single domain protein, which is a variant of the ubiquitous receiver domain (Pfam, response_reg; SMART, REC) that is found in response regulators of classic two-component signal transduction systems as well as chemotaxis systems (Galperin, 2006; West and Stock, 2001). In order to find stand-alone REC domains (CheY candidates), BLAST searches can be restricted to retrieve sequences of only a certain length or they can be identified by extensive domain architecture queries. Unfortunately, identifying CheY proteins in a set of stand-alone REC domains is still a serious challenge. In some two-component systems, stand-alone REC domains serve as middlemen in extended phosphotransfer relays (Hoch, 2000; Stock *et al.*, 2000).

Stand-alone REC domains that are more similar to experimentally characterized CheY proteins rather than components of phosphorelay systems are predicted to be CheY proteins. Gene neighborhood analysis is a powerful technique used to confidently identify CheY proteins because they are often encoded in a cluster of other chemotaxis genes. We can begin to identify CheY proteins computationally by building phylogenetic trees of stand-alone REC domains and searching for subfamilies that can be linked to chemotaxis by experimental evidence and gene neighborhood analysis. Once CheY proteins are predicted there is the added confusion of identifying what type of motility they regulate. Given that CheY proteins have been shown to regulate both flagellar and type IV pili-based motility, the whole-genome search for the presence or absence of genes encoding these motility organelles can aid in delineating functional subfamilies of CheY proteins.

CheB and CheR

The CheB methylesterase and CheR methyltransferase work together to regulate the methylation state of MCPs (Li and Hazelbauer, 2005). Although there are examples of flagellar and pili-based chemotaxis systems that lack CheB and CheR (Terry *et al.*, 2006; Whitchurch *et al.*, 2004; Zhulin, 2001), they are present in the vast majority of chemotaxis systems that have been studied experimentally or deduced from genome sequence. Unexpectedly, some chemotaxis systems that contain all core components may lack CheR but not CheB (Whitchurch *et al.*, 2004) or lack CheB but not CheR (e.g., the genomes of Listeria innocua, Listeria monocytogenes, Bacillus cereus, Bacillus anthracis, and *Bacillus thuringiensis*; K. Wuichet, unpublished observation). The genome of *Hyphomonas neptunium* that lacks all core chemotaxis components still contains the cheR gene, although no chemotaxis was detected in this organism (Badger *et al.*, 2006).

CheB is typically defined by the presence of the catalytic domain (Pfam, CheB_methylest) fused to a regulatory amino-terminal REC domain. In Pfam, CheR is defined by two domain models, CheR_N and CheR, whereas both of the regions corresponding to these domains are encompassed in the SMART MeTrc domain model (Fig. 1). The Pfam domains are a better reflection of the three-dimensional structure of the protein, which consists of two globular subdomains corresponding to the domain models; however, the N-terminal subdomain (CheR_N) is not highly conserved, which has resulted in a poor domain model. The CheR and MeTrc domains, which include the highly conserved catalytic region, are the best models to search for CheR proteins in genome databases.

Comparative genomics has already identified a variety of CheR domain architectures, including CheR fusions with C-terminal tetratricopeptide repeats (SMART, TPR) and N-terminal CheW domains (Shiomi *et al.*, 2002), as well as CheB association with class I histidine kinase domains (SMART: HisKA and Pfam: HWE_HK) (Karniol and Vierstra, 2004). TPR domains are known to promote protein-protein interactions (Lamb *et al.*, 1995). The CheR-TPR fusion proteins have been shown to be involved in chemotaxis, and their TPR domains are found to interact with their N-terminal CheR regions and with CheY (Bustamante *et al.*, 2004). A simple domain search for CheB reveals the histidine kinase fusion proteins along with stand-alone CheB catalytic domains that lack a regulatory REC domain. The roles of the adaptation domains of the kinase fusion proteins and the stand-alone CheB catalytic domains have not been determined experimentally. Comparative genomic analyses have the potential to reveal new insight into this adaptation system and identify specific targets for further experimental analysis.

CheC and CheX

The crystal structures of the closely related CheC and CheX proteins reveal distinct differences in their structures and interactions (Park *et al.*, 2004b). These two CheY phosphatases share sequence similarity, but have different structures and domain architectures. The CheC/CheX homolog FliY, a component of the flagellar motor, can be clearly discriminated from CheC and CheX by the presence of a C-terminal SpoA domain (Pfam, SpoA) that is involved in structural assembly. An exception from this rule is the split FliY protein of *T. maritima* (Park *et al.*, 2004b; Szurmant *et al.*, 2004). The CheC domain model (Pfam, CheC) is built from the active site of the enzyme. CheC and FliY have two homologous active sites (Fig. 1). The CheX phosphatase is closely related to CheC, but has only the second active site of CheC (Fig. 1). CheX was found to act as a dimer in the crystal structure, unlike CheC. CheX also differs from CheC in its length and secondary structure (Park *et al.*, 2004b). CheC has been cocrystallized with CheD, an interaction that has been shown to increase the phosphatase activity of CheC (Chao *et al.*, 2006). CheC and CheX are poorly conserved even at the very small active site region, which makes their identification by domain models rather difficult. Similarity searches, such as BLAST, are better suited for finding CheC and CheX homologs, although these searches are unable to discriminate between CheC and CheX (Fig. 8). Multiple sequence alignments are needed to confirm the validity of CheC/CheX/FliY protein family members identified by similarity searches. The VISSA program (Ulrich and Zhulin, 2005) can aid in distinguishing CheC and CheX given the distinct differences in their secondary structures. We find new CheC/CheX protein domain architectures including a fusion with CheA, fusions via duplication, and fusions with REC domains in a variety of Proteobacteria. Given that CheC and CheX act to dephosphorylate the CheY REC domain, it is possible that the REC-CheC/CheX fusion proteins promote dephosphorylation of the REC domain in the fusion. Experimental analysis is needed to clarify the function of these fusion proteins.

CheC is often encoded in the genome near CheD and/or a CheY-like protein, whereas CheX is typically not encoded near chemotaxis components with the exception of Spirochetes. CheC is also found in some genomes that lack CheD, for example, *Vibrio parahaemolyticus* and *Myxococcus xanthus*. The CheX protein has been shown to dephosphorylate CheY-P (Motaleb *et al.*, 2005; Park *et al.*, 2004b) and interact with CheA (Sim *et al.*, 2005); however, it remains to be seen whether CheX acts as a phosphatase and plays a role in chemotaxis in more distantly related organisms.

CheD

In addition to playing a role in the excitation pathway by aiding CheY-P dephosphorylation by CheC, CheD also plays a role in the adaptation pathway by deamidating key glutamine residues of MCPs into glutamate residues so they can be methylated by CheR (Kristich and Ordal, 2002). Similarity searches reveal that CheD is highly conserved and can be easily identified solely by queries for its domain model (Pfam, CheD). The phyletic distribution of CheD and CheC showed that many organisms that have CheD lack CheC (Kirby *et al.*, 2001). CheD is a single domain protein, but its fusion with CheB can be seen in *Bdellovibrio bacteriovorans*. Our gene neighborhood analysis showed that the overwhelming majority of CheD proteins are encoded in the genomes near other chemotaxis proteins, implicating their involvement in chemotaxis regardless of the presence of CheC.

CheZ

Although the CheZ phosphatase of CheY was previously found only in some representatives of *β*/*γ*-Proteobacteria (Szurmant and Ordal, 2004), experiments have identified a divergent CheZ the protein, which was not detected by current Pfam domain model (Pfam, CheZ) in the member of *ε*-Proteobacteria, *Helicobacter pylori* (Terry *et al.*, 2006). We performed PSI-BLAST searches against completely sequenced genomes to identify many other previously undetected members of the CheZ family from different species, including representatives of *α*- and *δ*-Proteobacteria (Fig. 9).

A multiple alignment reveals that all of the sequences identified in *α*- and *δ*-Proteobacteria form a specific CheZ subfamily, which can be distinguished by the conserved catalytic glutamine residue and high conservation of positions surrounding the catalytic residue. The phylogenetic tree built from the multiple alignment suggests three subfamilies of CheZ proteins based on sequence features and taxonomy (Fig. 9). Although CheZ has been shown to interact with both CheY and CheA, the subfamily of the *α*- and *δ*-proteobacterial sequences lacks the CheA-binding region entirely, and thus these CheZ proteins are predicted not to interact with CheA. None of these proteins have been experimentally characterized, but the CheZ of *Caulobacter crescentus* is located near a chemotaxis locus containing *cheA, cheB, cheR, cheW*, and *CheY*, which supports the hypothesis that representatives of this subfamily play a role in chemotaxis. The experimentally characterized CheZ of *E. coli* is found in the *β*/*γ*-Proteobacteria subfamily. The CheZ of *E. coli* interacts with CheY-P at two distinct regions (Zhao *et al.*, 2002), and a third region interacts with CheA (Cantwell *et al.*, 2003). Both the CheY and the CheA interaction regions are found in all members of the *β*/*γ* subfamily except *Xanthomonas axonopodis* and *Xanthomonas campestris*, which lack the CheA-binding region. The *ε*-Proteobacteria subfamily has an elongated CheA-binding region that shares no sequence similarity with the CheA-binding region of the *β*/*γ* subfamily, but the presence of this subdomain suggests that it may still be involved in binding CheA. The CheZ of *H. pylori* has been implicated in chemotaxis, but direct interaction studies have not been carried out (Terry *et al.*, 2006). Although the *ε*- and *α*/*δ*-proteobacterial subfamilies are quite divergent, the conservation of catalytic residues suggests that they are involved in dephosphorylation of CheY-P proteins, and the phyletic distribution suggests that CheZ originated in a common ancestor of Proteobacteria.

CheW and CheV

CheW and CheV have both been shown to be involved in sensory lattice scaffolding by interacting with CheA and MCPs (Gegner *et al.*, 1992; Rosario *et al.*, 1994). As seen in Fig. 1, the CheW protein is a single domain (Pfam, CheW; SMART, CheW), but domain queries with CheW will identify multiple components of the chemotaxis system, as it is homologous to domains founds in all CheV and CheA proteins in addition to an unusual CheW-CheR fusion protein found exclusively in Spirochetes. Searches that include the CheW domain while excluding REC, HATPase_c and MeTrc domains should identify true CheW proteins. Phyletic distribution shows that the CheW protein is essential to all chemotaxis systems with the exceptions of L. innocua, L. monocytogenes, B. cereus, B. anthracis, and *B. thuringiensis*, which appear to exclusively use CheV in place of CheW. CheW proteins are typically found in major chemotaxis loci together with CheA. CheW is a subject of frequent domain duplication events. On a phylogenetic tree the duplicate CheWs often fall into different subgroups, which raise questions as to the function of these multiple CheW proteins. There are also a few proteins that contain multiple CheW domains that are significantly diverged in sequence. The lack of distinct subfamilies that can be grouped by consistent gene neighborhoods shows that more detailed sequence and structure analysis is needed to derive meaningful conclusions about the functional implications of CheW diversity. One attractive hypothesis is that various CheW paralogs recognize specific classes of MCP signaling domain and thus link particular MCPs to individual signal transduction pathways.

The CheV protein is composed of a CheW domain and a C-terminal REC domain. The CheA kinase regulates the phosphorylation state of the REC domain in order to modulate CheV function (Karatan *et al.*, 2001). Domain searches for proteins that have CheW and REC domains, but not HATPase_c domains, should clearly identify CheV proteins. The phyletic distribution shows that CheV are only present in a few representatives of Firmicutes and ε, δ, and β/γ classes of Proteobacteria. The disparate distribution of CheV does not allow us to clearly delineate its evolutionary origins. Specific subfamilies and duplication events can be identified from a phylogenetic tree, but there are no sequence features that can be related to functional differences among the subfamilies. CheV proteins are sometimes found near flagellar proteins and CheR, and rarely near CheA, but they are not typically encoded near other chemotaxis proteins.

Because both CheW and CheV have primary roles in scaffolding with fewer dynamic interactions than the other chemotaxis proteins, it is possible that the evolutionary pressures on these components have resulted in more divergence at the individual sequence level rather than easily identifiable insertion and deletion events. This is supported by the observation that such divergence in sequence between distant CheW homologs does not prevent functional complementation (Alexandre and Zhulin, 2003).

References

Acuna G, Shi W, Trudeau K, Zusman DR. The ‘CheA’ and ‘CheY’ domains of *Myxococcus xanthus* FrzE function independently *in vitro* as an autokinase and a phosphate acceptor, respectively. FEBS Lett 1995;358:31–33. [PubMed: 7821424]

Alexandre G, Zhulin IB. Different evolutionary constraints on chemotaxis proteins CheW and CheY revealed by heterologous expression and protein sequence analysis. J. Bacteriol 2003;185:544–552. [PubMed: 12511501]

Alexander, RP.; Zhulin, IB. Submitted for publication. 2007.

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. [PubMed: 9254694]

Anantharaman V, Aravind L. Cache: A signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors. Trends Biochem. Sci 2000;25:535–537. [PubMed: 11084361]

Aravind L, Ponting CP. The GAF domain: An evolutionary link between diverse phototransducing proteins. Trends Biochem. Sci 1997;22:458–459. [PubMed: 9433123]

Aravind L, Ponting CP. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signaling proteins. FEMS Microbiol Lett 1999;176:111–116. [PubMed: 10418137]

Badger JH, Hoover TR, Brun YV, Weiner RM, Laub MT, Alexandre G, Mrazek J, Ren Q, Paulsen IT, Nelson KE, Khouri HM, Radune D, et al. Comparative genomic evidence for a close relationship between the dimorphic prosthecate bacteria *Hyphomonas neptunium* and *Caulobacter crescentus*. J. Bacteriol 2006;188:6841–6850. [PubMed: 16980487]

Baker MD, Wolanin PM, Stock JB. Signal transduction in bacterial chemotaxis. Bioessays 2006;28:9–22. [PubMed: 16369945]

Berleman JE, Bauer CE. Involvement of a Che-like signal transduction cascade in regulating cyst cell development in *Rhodospirillum centenum*. Mol. Microbiol 2005;56:1457–1466. [PubMed: 15916598]

Bhaya D, Takahashi A, Grossman AR. Light regulation of type IV pilus-dependent motility by chemosensor-like elements in *Synechocystis* PCC6803. Proc. Natl. Acad. Sci. USA 2001;98:7540–7545. [PubMed: 11404477]

Bibikov SI, Biran R, Rudd KE, Parkinson JS. A signal transducer for aerotaxis in *Escherichia coli*. J. Bacteriol 1997;179:4075–4079. [PubMed: 9190831]

Bilwes AM, Alex LA, Crane BR, Simon MI. Structure of CheA, a signal-transducing histidine kinase. Cell 1999;96:131–141. [PubMed: 9989504]

Brooun A, Bell J, Freitas T, Larsen RW, Alam M. An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins. J. Bacteriol 1998;180:1642–1646. [PubMed: 9537358]

Bustamante VH, Martinez-Flores J, Vlamakis HC, Zusman DR. Analysis of the Frz signal transduction system of *Myxococcus xanthus* shows the importance of the conserved C-terminal region of the cytoplasmic chemoreceptor FrzCD in sensing signals. Mol. Microbiol 2004;53:1501–1513. [PubMed: 15387825]

Cantwell BJ, Draheim RR, Weart RB, Nguyen C, Stewart RC, Manson MD. CheZ phosphatase localizes to chemoreceptor patches via CheA-short. J. Bacteriol 2003;185:2354–2361. [PubMed: 12644507]

Chao X, Muff TJ, Park SY, Zhang S, Pollard AM, Ordal GW, Bilwes AM, Crane BR. A receptor-modifying deamidase in complex with a signaling phosphatase reveals reciprocal regulation. Cell 2006;124:561–571. [PubMed: 16469702]

Chenna R, Sugawara H, Koke T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31:3497–3500. [PubMed: 12824352]

Crooks EG, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res 2004;14:1188–1190. [PubMed: 15173120]

Cserzo M, Eisenhaber F, Eisenhaber B, Simon I. On filtering false positive transmembrane protein predictions. Protein Eng 2002;15:745–752. [PubMed: 12456873]

Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: The dense alignment surface method. Protein Eng 1997;10:673–676. [PubMed: 9278280]

D'Argenio DA, Calfee MW, Rainey PB, Pesci EC. Autolysis and autoaggregation in *Pseudomonas aeruginosa* colony morphology mutants. J. Bacteriol 2002;184:6481–6489. [PubMed: 12426335]

Dutta R, Qin L, Inouye M. Histidine kinases: Diversity of domain organization. Mol. Microbiol 1999;34:633–640. [PubMed: 10564504]

Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797. [PubMed: 15034147]

Felsenstein J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 1989;5:164–166.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, et al. Pfam: Clans, web tools and services. Nucleic Acids Res 2006;34:D247–D251. [PubMed: 16381856]

Galperin MY. Structural classification of bacterial response regulators: Diversity of output domains and domain combinations. J. Bacteriol 2006;188:4169–4182. [PubMed: 16740923]

Galtier N, Gouy M, Gautier G. SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci 1996;12:543–548. [PubMed: 9021275]

Gegner JA, Graham DR, Roth AF, Dahlquist FW. Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway. Cell 1992;70:975–982. [PubMed: 1326408]

Guvener ZT, Tifrea DF, Harwood CS. Two different *Pseudomonas aeruginosa* chemosensory signal transduction complexes localize to cell poles and form and remold in stationary phase. Mol. Microbiol 2006;61:106–118. [PubMed: 16824098]

Hickman JW, Tifrea DF, Harwood CS. A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. Proc. Natl. Acad. Sci. USA 2005;102:14422–14427. [PubMed: 16186483]

Hoch JA. Two-component and phosphorelay signal transduction. Curr. Opin. Microbiol 2000;3:165–170. [PubMed: 10745001]

Hou S, Larsen RW, Boudko D, Riley CW, Karatan E, Zimmer M, Ordal GW, Alam M. Myoglobin-like aerotaxis transducers in Archaea and Bacteria. Nature 2000;403:540–544. [PubMed: 10676961]

Hulko M, Berndt F, Gruber M, Linder JU, Truffault V, Schultz A, Martin J, Schultz JE, Lupas AN, Coles M. The HAMP domain structure implies helix rotation in transmembrane signaling. Cell 2006;126:929–940. [PubMed: 16959572]

Karatan E, Saulmon MM, Bunn MW, Ordal GW. Phosphorylation of the response regulator CheV is required for adaptation to attractants during *Bacillus subtilis* chemotaxis. J. Biol. Chem 2001;276:43618–43626. [PubMed: 11553614]

Karniol B, Vierstra RD. The HWE histidine kinases, a new family of bacterial two-component sensor kinases with potentially diverse roles in environmental signaling. J. Bacteriol 2004;186:445–453. [PubMed: 14702314]

Kato M, Mizuno T, Shimizu T, Hakoshima T. Insights into multistep phosphorelay from the crystal structure of the C-terminal HPt domain of ArcB. Cell 1997;88:717–723. [PubMed: 9054511]

Kirby JR, Kristich CJ, Saulmon MM, Zimmer MA, Garrity LF, Zhulin IB, Ordal GW. CheC is related to the family of flagellar switch proteins and acts independently from CheD to control chemotaxis in *Bacillus subtilis*. Mol. Microbiol 2001;42:573–585. [PubMed: 11722727]

Kirby JR, Zusman DR. Chemosensory regulation of developmental gene expression in *Myxococcus xanthus*. Proc. Natl. Acad. Sci. USA 2003;100:2008–2013. [PubMed: 12566562]

Kristich CJ, Ordal GW. *Bacillus subtilis* CheD is a chemoreceptor modification enzyme required for chemotaxis. J. Biol. Chem 2002;277:25356–25362. [PubMed: 12011078]

Kumar S, Tamura K, Nei M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinformatics 2004;5:150–163. [PubMed: 15260895]

Lamb JR, Tugendreich S, Hieter P. Tetratrico peptide repeat interactions: To TPR or not to TPR? Trends Biochem. Sci 1995;20:257–259. [PubMed: 7667876]

LeMoual H, Koshland DE. Molecular evolution of the C-terminal cytoplasmic domain of a superfamily of bacterial receptors involved in taxis. J. Mol. Biol 1996;261:568–585. [PubMed: 8794877]

Li MS, Hazelbauer GL. Adaptational assistance in clusters of bacterial chemoreceptors. Mol. Microbiol 2005;56:1617–1626. [PubMed: 15916610]

Martin AC, Wadhams GH, Armitage JP. The roles of the multiple CheW and CheA homologues in chemotaxis and in chemoreceptor localization in *Rhodobacter sphaeroides*. Mol. Microbiol 2001;40:1261–1272. [PubMed: 11442826]

McEvoy MM, Hausrath AC, Randolph GB, Remington SJ, Dahlquist RM. Two binding modes reveal flexibility in kinase/response regulator interactions in the bacterial chemotaxis pathway. Proc. Natl. Acad. Sci. USA 1998;95:7333–7338. [PubMed: 9636149]

McGinnis S, Madden TL. BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004;32:W20–W25. [PubMed: 15215342]

Motaleb MA, Miller MR, Li C, Bakker RG, Goldstein SF, Silversmith RE, Bourret RB, Charon NW. CheX is a phosphorylated CheY phosphatase essential for *Borrelia burgdorferi* chemotaxis. J. Bacteriol 2005;187:7963–7969. [PubMed: 16291669]

Notredame C, Higgins DG, Heninga J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol 2000;302:205–217. [PubMed: 10964570]

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA 1999;96:2896–2901. [PubMed: 10077608]

Park SY, Beel BD, Simon MI, Bilwes AM, Crane RB. In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. Proc. Natl. Acad. Sci. USA 2004a; 101:11646–11651. [PubMed: 15289606]

Park SY, Chao X, Gonzalez-Bonet G, Beel BD, Bilwes AM, Crane BR. Structure and function of an unusual family of protein phosphatases: The bacterial chemotaxis proteins CheC and CheX. Mol. Cell 2004b;16:563–574. [PubMed: 15546616]

Pei JM, Sadreyev R, Grishin NV. PCMA: Fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 2003;19:427–428. [PubMed: 12584134]

Pittman MS, Goodwin M, Kelly DJ. Chemotaxis in the human gastric pathogen *Helicobacter pylori*: Different roles for CheW and the three CheV paralogues, and evidence for CheV2 phosphorylation. Microbiology 2001;147:2493–2504. [PubMed: 11535789]

Porter SL, Armitage JP. Chemotaxis in *Rhodobacter sphaeroides* requires an atypical histidine protein kinase. J. Biol. Chem 2004;279:54573–54580. [PubMed: 15485885]

Porter SL, Warren AV, Martin AC, Armitage JP. The third chemotaxis locus of *Rhodobacter sphaeroides* is essential for chemotaxis. Mol. Microbiol 2002;46:1081–1094. [PubMed: 12421313]

Rebbapragada A, Johnson MS, Harding GP, Zuccarelli AJ, Fletcher HM, Zhulin IB, Taylor BL. The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior. Proc. Natl. Acad. Sci. USA 1997;94:10541–10546. [PubMed: 9380671]

Reinelt S, Hofmann E, Gerharz T, Bott M, Madden DR. The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain. J. Biol. Chem 2003;278:39189–39196. [PubMed: 12867417]

Rosario MM, Fredrick KL, Ordal GW, Helmann JD. Chemotaxis in *Bacillus subtilis* requires either of two functionally redundant CheW homologs. J. Bacteriol 1994;176:2736–2739. [PubMed: 8169224]

Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–3385. [PubMed: 12824332]

Shiomi D, Zhulin IB, Homma M, Kawagishi I. Dual recognition of the bacterial chemoreceptor by chemotaxis-specific domains of the CheR methyltransferase. J. Biol. Chem 2002;277:42325–42333. [PubMed: 12101179]

Shu CJ, Ulrich LE, Zhulin IB. The NIT domain: A predicted nitrate-responsive module in bacterial sensory receptors. Trends Biochem. Sci 2003;28:121–124. [PubMed: 12633990]

Sim JH, Shi W, Lux R. Protein-protein interactions in the chernotaxis signaling pathway of *Treponema denticola*. Microbiology 2005;151:1801–1807. [PubMed: 15941989]

Stock AM, Robinson VL, Goudreau PN. Two-component signal transduction. Annu. Rev. Biochem 2000;69:183–215. [PubMed: 10966457]

Sun H, Zusman DR, Shi W. Type IV pilus of *Myxococcus xanthus* is a motility apparatus controlled by the frz chemosensory system. Curr. Biol 2000;10:1143–1146. [PubMed: 10996798]

Szurmant H, Muff TJ, Ordal GW. *Bacillus subtilis* CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade. J. Biol. Chem 2004;279:21787–21792. [PubMed: 14749334]

Szurmant L, Ordal GW. Diversity in chemotaxis mechanisms among the bacteria and archaea. Microbiol. Mol. Biol. Rev 2004;68:301–319. [PubMed: 15187186]

Taylor BL, Zhulin IB. PAS domains: Internal sensors of oxygen, redox potential and light. Microbiol. Mol. Biol. Rev 1999;63:479–506. [PubMed: 10357859]

Terry K, Go AC, Ottemann KM. mapping of Proteomic a suppressor of non-chemotactic *cheW* mutants reveals that *Helicobacter pylori* contains a new chemotaxis protein. Mol. Microbiol 2006;61:871–882. [PubMed: 16879644]

Ulrich LE, Koonin EV, Zhulin IB. One-component systems dominate signal transduction in prokaryotes. Trends Microbiol 2005;13:52–56. [PubMed: 15680762]

Ulrich LE, Zhulin IB. Four-helix bundle: A ubiquitous sensory module in prokaryotic signal transduction. Bioinformatics 2005;21(Suppl 3):iii45–iii48. [PubMed: 16306392]

Ulrich IE, Zhulin IB. MiST: The Microbial Signal Transduction database. Nucleic Acids Res 2007;35

Wadhams GH, Armitage JP. Making sense of it all: Bacterial chemotaxis. Nat. Rev. Mol. Cell Biol 2004;5:1024–1037. [PubMed: 15573139]

West AH, Stock AM. Histidine kinases and response regulator proteins in two-component signaling systems. Trends Biochem. Sci 2001;26:369–376. [PubMed: 11406410]

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2006;34:D173–D180. [PubMed: 16381840]

Whitchurch CB, Leech AJ, Young MD, Kennedy D, Sargent JL, Bertrand JJ, Semmler AB, Mellick AS, Martin PR, Alm RA, Hobbs M, Beatson SA, et al. Characterization of a complex chemosensory signal transduction system which controls twitching motility in *Pseudomonas aeruginosa*. Mol. Microbiol 2004;52:873–893. [PubMed: 15101991]

Williams SB, Stewart V. Functional similarities among two-component sensors and methyl-accepting chemotaxis proteins suggest a role for linker region amphipathic helices in transmembrane signal transduction. Mol. Microbiol 1999;33:1093–1102. [PubMed: 10510225]

Wuichet K, Zhulin IB. Molecular evolution of sensory domains in cyanobacterial chemoreceptors. Trends Microbiol 2003;11:200–203. [PubMed: 12781518]

Yoshihara S, Geng X, Ikeuchi M. pilG Gene cluster and split pilL genes involved in pilus biogenesis, motility and genetic transformation in the cyanobacterium *Synechocystis* sp. PCC 6803. Plant Cell Physiol 2002;43:513–521. [PubMed: 12040098]

Zhang W, Phillips GN. Structure of the oxygen sensor in *Bacillus subtilis*: Signal transduction of chemotaxis by control of symmetry. Structure 2003;11:1097–1110. [PubMed: 12962628]

Zhao R, Collins EJ, Bourret RB, Silversmith RE. Structure and catalytic mechanism of the *E. coli* chemotaxis phosphatase CheZ. Nat. Struct. Biol 2002;9:570–575. [PubMed: 12080332]

Zhulin IB. The superfamily of chemotaxis transducers: From physiology to genomics and back. Adv. Microb. Physiol 2001;45:157–198. [PubMed: 11450109]

Zhulin IB, Nikolskaya AN, Galperin MY. Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea. J. Bacteriol 2003;185:285–294. [PubMed: 12486065]

Further Reading

Barnakov AN, Barnakova LA, Hazelbauer GL. Location of the receptor-interaction site on CheB, the methylesterase response regulator of bacterial chemotaxis. J. Biol. Chem 2001;276:32984–32989. [PubMed: 11435446]

Djordjevic S, Stock AM. Chemotaxis receptor recognition by protein methyl-transferase CheR. Nat. Struct. Biol 1998;5:446–450. [PubMed: 9628482]

Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol 2004;338:1027–1036. [PubMed: 15111065]

Lai WC, Hazelbauer GL. Carboxyl-terminal extensions beyond the conserved pentapeptide reduce rates of chemoreceptor adaptational modification. J. Bacteriol 2005;187:5115–5121. [PubMed: 16030204]

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5.0: Domains in the context of genomes and networks. Nucleic Acids Res 2006;32:D142–D144. [PubMed: 14681379]

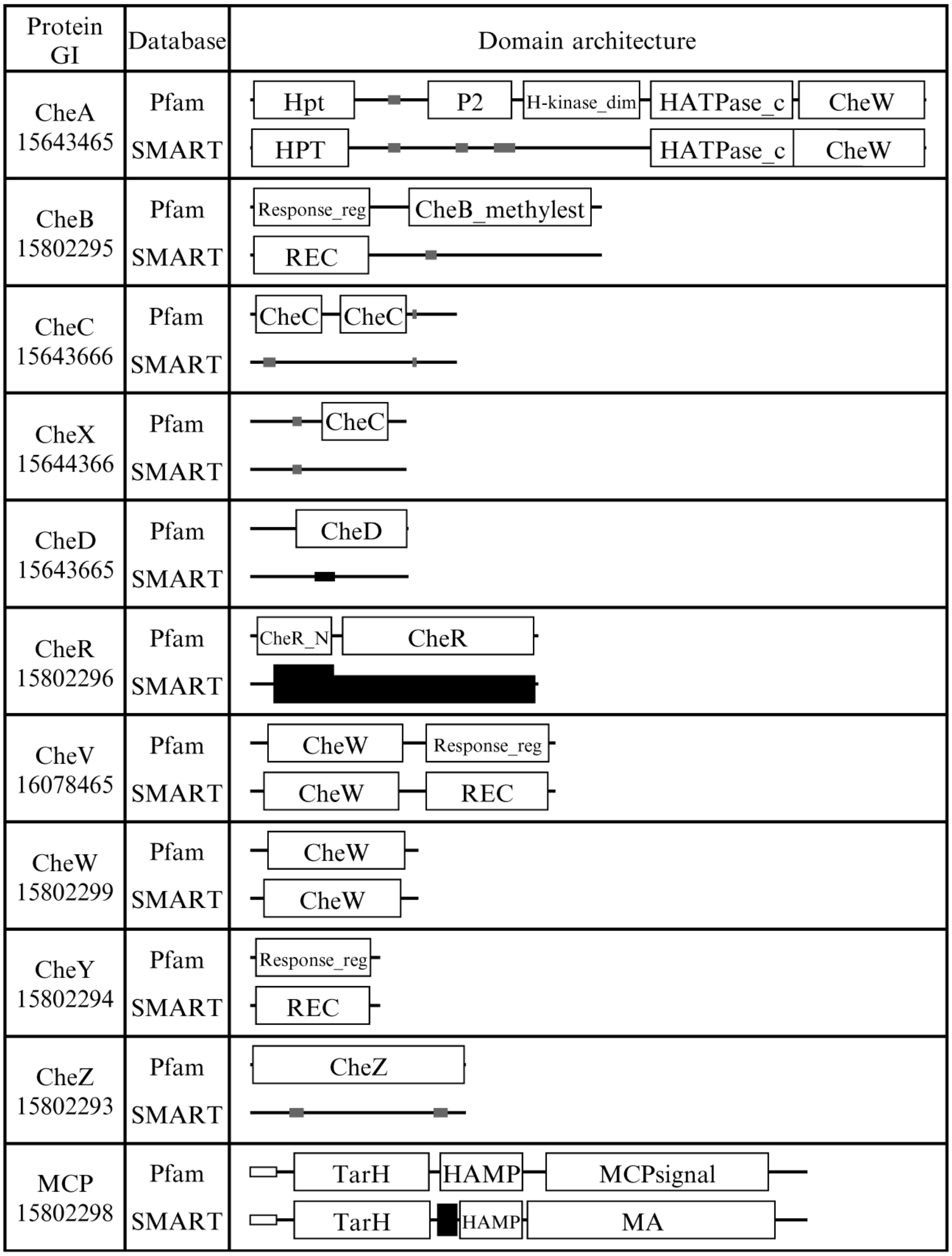

FIG. 1.

Domain architecture of chemotaxis proteins as visualized in MiST. The MiST database (Ulrich and Zhulin, 2007) uses the domain models from both Pfam and SMART databases. Domains are shown as white boxes with their names inside. Small black, gray, and white boxes indicate predicted transmembrane, low complexity, and signal peptide regions, respectively. The NCBI database GI (GenBank identifier) numbers corresponding to each protein sequence are given under their respective protein names.

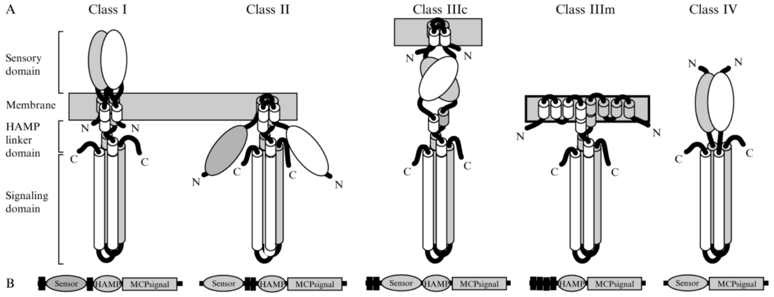

FIG. 2.

MCP membrane topology classes. Differing membrane topology divides MCPs into four main classes. (A) Schematic representation of the three-dimensional structure of MCP dimers of different sensor classes. Oval domains are sensory domains of varied secondary structure. Cylinders represent *α*-helical and coiled coil regions. MCP monomers are differentiated by gray and white coloring. Class I, transmembrane MCPs with extracellular sensory domains; class II, membrane-bound MCPs with N-terminal cytoplasmic sensory domains; class III, membrane-bound MCPs with cytoplasmic sensory domains located C-terminally to the last transmembrane regions (IIIc) or without sensory domains (IIIm); class IV, cytoplasmic MCPs. (B) MCP sensor class can be determined from domain architecture where transmembrane regions and domains are well predicted. Transmembrane regions are indicated by black boxes

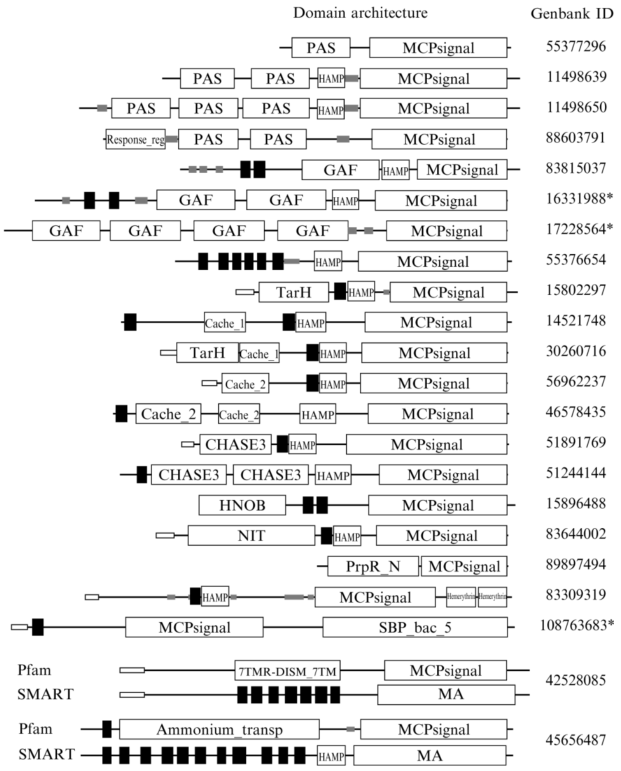
Domain architecture
Genbank ID
PAS MCPsignal 55377296
PAS PAS HAMP MCPsignal 11498639
PAS PAS PAS HAMP MCPsignal 11498650
Response_reg PAS PAS MCPsignal 88603791
GAF HAMP MCPsignal 83815037
GAF GAF HAMP MCPsignal 16331988*
GAF GAF GAF GAF MCPsignal 17228564*
HAMP MCPsignal 55376654
TarH HAMP MCPsignal 15802297
Cache_1 HAMP MCPsignal 14521748
TarH Cache_1 HAMP MCPsignal 30260716
Cache_2 HAMP MCPsignal 56962237
Cache_2 Cache_2 HAMP MCPsignal 46578435
CHASE3 HAMP MCPsignal 51891769
CHASE3 CHASE3 HAMP MCPsignal 51244144
HNOB MCPsignal 15896488
NIT HAMP MCPsignal 83644002
PrpR_N MCPsignal 89897494
HAMP MCPsignal Hemerythrin Hemerythrin 83309319
MCPsignal SBP_bac_5 108763683*
Pfam 7TMR-DISM_7TM MCPsignal
SMART MA
42528085
Pfam Ammonium_transp MCPsignal
SMART HAMP MA
45656487

FIG. 3.

Diversity of sensory domains in MCPs. All sensory domains are Pfam domain models, except the GAF domain, which is the SMART model (it is slightly longer than the Pfam domain model). HAMP domains are the SMART domain model. MCPs containing hemerythrin and SBP_bac_5 sensory domains represent the atypical topology where the MCP signaling domain is N-terminal of the sensory domain. The Pfam TarH model has shown to be erroneous and will soon be replaced by a correct model termed 4HB_MCP (Ulrich and Zhulin, 2005). Both Pfam and SMART domain architectures are shown for two MCPs with class IIIm membrane topology. Small gray and white boxes indicate predicted low complexity and signal peptide regions, respectively. Black boxes represent transmembrane regions. Long sequences marked by an asterisk (*) were shortened for display and are not to scale.

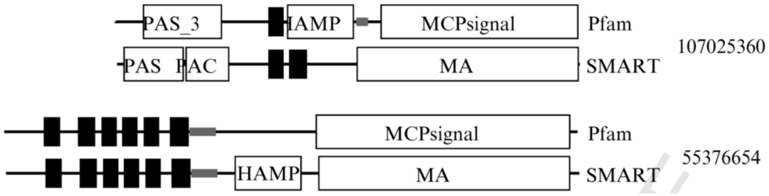

FIG. 4.

HAMP domain models are imperfect. Both the Pfam SMART HAMP domain models have low sensitivity; however, implementation of both models in MiST enables the identification of HAMP domains in many cases when one of the domain database models misses the target. Note that the Pfam HAMP domain models often (but not always) overlap with one of the transmembrane regions.

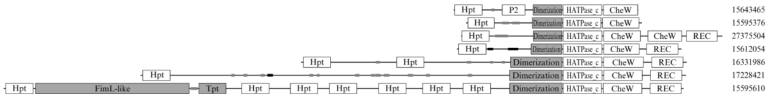

FIG. 5.

A common core and diversity of CheA homologs. The domain architectures of selected CheA proteins are shown with their corresponding NCBI GI numbers to the right. All shown domains are from Pfam except for the REC domain (SMART domain model). The dimerization domains shown in gray were delineated by PSI-BLAST analysis; current dimerization domain models have very low sensitivity and fail to predict the domain in many instances. Our analysis shows that the dimerization domain is present in all CheA homologs identified to date (K. Wuichet, unpublished data). Small black, gray, and white boxes indicate predicted transmembrane, low complexity, and signal peptide regions, respectively. The FimL-like domain shows similarity to the FimL pili motility protein, and the Tpt domain shows similarity to Hpt domains, but it has a threonine in place of the conserved histidine (the phosphorylation site). Despite diverse domain architectures, all CheA proteins contain Hpt, dimerization, HATPase_c, and CheW domains, with the latter three forming in a tight protein core. CheA-CheC fusion proteins were also identified; see Fig. 8.

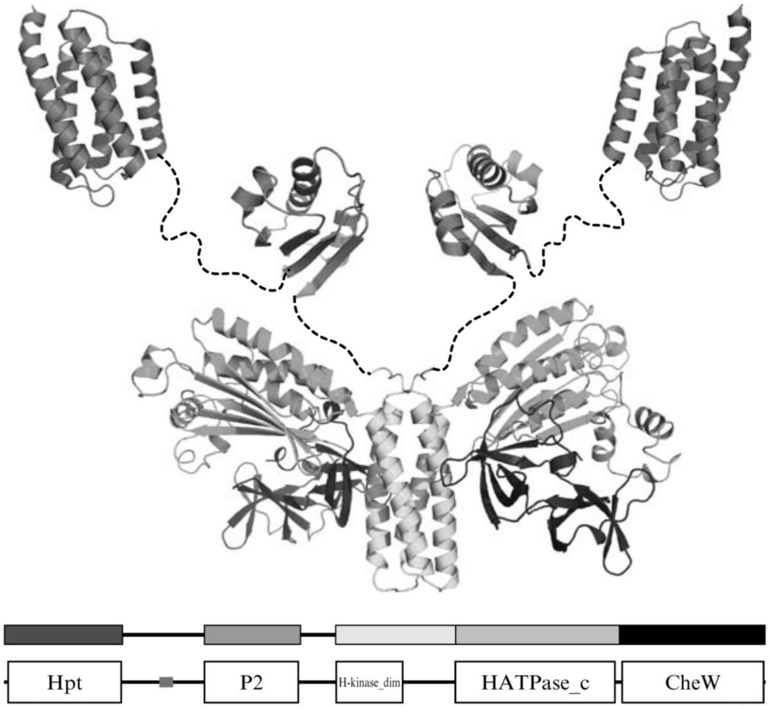

FIG. 6.

The relationship between the domain architecture and the structure of CheA. The domain architecture of the CheA protein directly relates to its structure. The Pfam domain model of CheA (GI 15643465) and its two-dimensional color scheme are shown below the three-dimensional model that has a matching color code. The three-dimensional model consists of three different crystal structures: the Hpt (or P1) domain (PDB identifier 1I5N), the P2 domain (1UOS), and the three core domains (PDB, 1BDJ)—dimerization (or P3) (Pfam, H-kinase_dim), HATPase_c (or P4), and CheW (or P5), respectively, with the linker regions hand drawn. The first two linker regions found in the domain architecture are predicted to be loops between the globular Hpt and P2 domains. The third predicted linker region of CheA suggests that the H-kinase_dim domain model does not capture the entire dimerization domain.

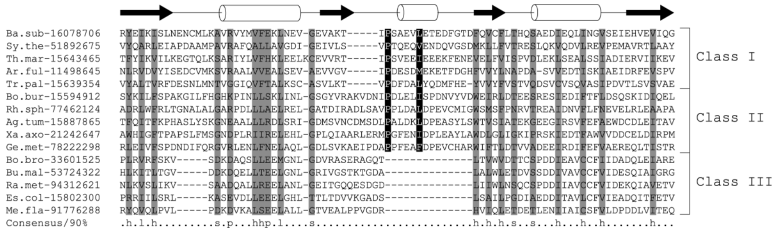
Ba.sub-16078706 RYEIKISLNENCMLKAVRVYMVFEKLNEV-GEVAKT-----IPSAEVLETEDFGTDFQVCFLTHQSAEDIEQLINGVSEIEHVEVIQG
Sy.the-51892675 VYQARLEIAPDAAMPAVRAFQALLAVGDI-GEIVLS-----VPTQEQVENDQVGSDMKLLFVTRESLQKVQDVLREVPEMAVRTLAAY
Th.mar-15643465 TFYIKVILKEGTQLKSARIYLVFHKLEELKCEVVRT-----IPSVEEIEEEKFENEVELFVISPVDLEKLSEALSSIADIERVIIKEV
Ar.ful-11498645 NLRVDVYISEDCVMKSVRAALVVEALSEV-AEVVGV-----IPDESDMEKETFDGHFVVYLNAPDA-SVVEDTISKIAEIDRFEVSPV
Tr.pal-15639354 VYALTVRFDESNLMNTVGGIQVFTALKSC-ASVLKT-----VPDFDALYQDMFHE-YVVYFVSTVQDSVCVSQVASIPDVTLSVSVAE
Bo.bur-15594912 SYKILFSPAKGILFHGHKPINLLSKLINL-GSGYVRAKVDNIPDLELISPDNVYVDWEIRLDTEESRESIEDIFTFLDSQSKIDIQEL
Rh.sph-77462124 ADRLWFRLTGNALALGARPDLLLAELREL-GATDIRADLSAVPPLDALDPEVCMIGWSMSFPNRVTREAIDNVFLFNEVELRLEAAPA
Ag.tum-15887865 TFQITFKPHASLYSKGNEAALLLRDLSRI-GDMSVNCDMSDLPALDKLDPEASYLSWTVSIATEKGEEGIRSVFEFAEWDCDLEITAV
Xa.axo-21242647 AWHIGFTPAPSLFMSGNDPLRIIRELEHL-GPLQIAARLERMPGFENIDPLEAYLAWDLGLIGKIPRSKIEDTFAWVVDDCELDIRPM
Ge.met-78222298 RLEIVFSPDNDIFQRGVRLENLFNELAQL-GDLSVKAEIPDAPPFEAFDPEVCHARWIFTLDTVVADEEIRDIFEFVAEREQLTISTR
Bo.bro-33601525 PLRVRFSKV-----SDKDAQSLLEEMGNL-GDVRASERAGQT--------------LTVWVDTTCSPDDIEAVCCFIIDADQLEIARE
Bu.mal-53724322 HLKITLTGV-----DDKDQELLTEELGNL-GRIVGSTKTGDA--------------LALWLETDVSSDDIVAVCCFVIDESQIAIGRG
Ra.met-94312621 NLKVSLIKV-----SAADQALLREELANL-GEITGQQESDGD--------------LIIWLNSQCSPDDIIAVCCFVIDEKQIAVETV
Es.col-15802300 PRRIILSRL-----KASEVDLLEEELGHL-TTLTDVVKGADS--------------LSAILPGDIAEDDITAVLCFVIEADQITFETV
Me.fla-91776288 RYQVQLPVL-----PDKDVKALSEELALL-GTVEALPPVGDR--------------HVIQLETDETLENIIAICSFVLDPDDLVITEQ
Consensus/90% .h.l.h.........s.p...hhp.l....s.........................h.h.h.s..s...h..h..h.......h...
Class I
Class II
Class III

FIG. 7.

Multiple alignment of the P2 domain and its classification. Three subclasses of the P2 domain were identified. A multiple alignment with representative members of each class of P2 domain shows the insertions and deletions that define each class. Positions conserved at 90% or more in an alignment of 116 P2 sequences are shown in gray. Conservation consensus is shown underneath the alignment (h, hydrophobic; l, aliphatic; p, polar; s, small). Black columns show conserved proline and hydrophobic positions in classes I and II. The secondary structure elements are shown above the alignment based on crystal structures from *E. coli* and *T. maritima* (McEvoy, 1998; Park, 2004a,b) Black arrows represent *β* strands. White cylinders represent *α* helices. Species abbreviations and NCBI GI numbers for each sequence are given at the left (full species name can be found by searching the NCBI nonredundant database with the corresponding GI number).

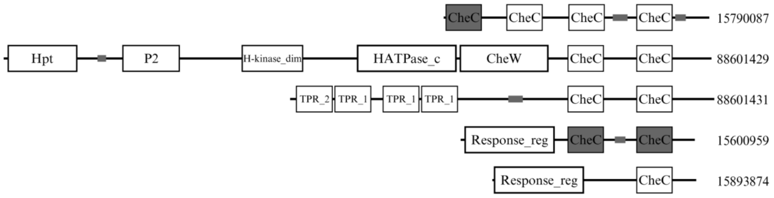

FIG. 8.

Diversity of CheC homologs. CheC and CheX proteins can be fused to different domains and proteins. Domains shown in gray were missed by the current domain models and were found by PSI-BLAST searches. Their approximate position in corresponding protein sequences is shown. Domain models are from Pfam. Small gray boxes indicate predicted low complexity regions. The NCBI GI number associated with each sequence is shown at the right.

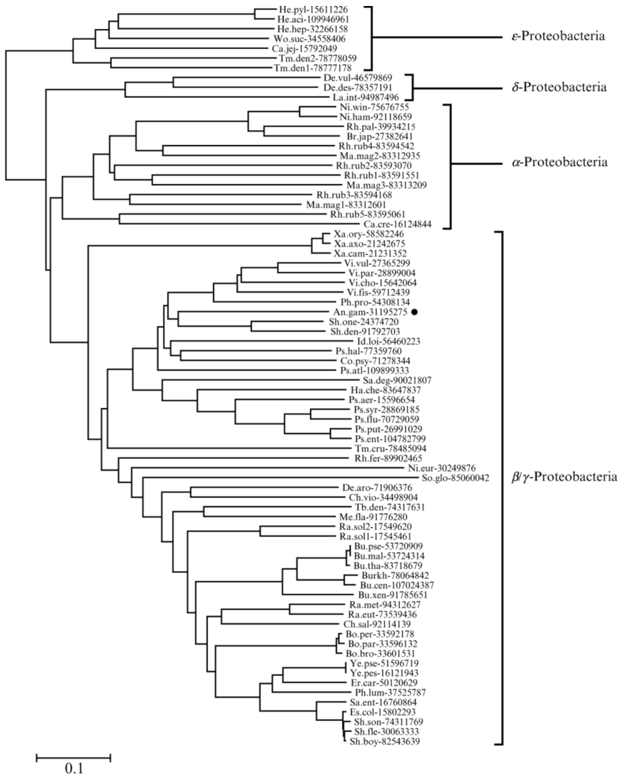

FIG. 9.

Neighbor-joining tree of the extended CheZ protein family. The CheZ protein family has members present in all classes of Proteobacteria, and the phylogenetic tree suggests its vertical evolution. The sequence identified by a black circle comes from a likely contamination with prokaryotic DNA in the genome of the mosquito *Anopheles gambiae*.