For submission to *Proteins: Structure, Function, and Bioinformatics*

3/2/22

Analysis of CheW-like domains provides insights into organization of prokaryotic chemotaxis systems

Luke R. Vass[1,#a], Robert B. Bourret[1*], Clay A. Foster[1,#b]

[1]Department of Microbiology & Immunology, University of North Carolina, Chapel Hill, North Carolina, United States of America

[#a]Current Address: Department of Pathology, University of Virginia, Charlottesville, Virginia, United States of America

[#b]Current Address: Department of Pediatrics, Section Hematology/Oncology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, United States of America

Running Title: CheW-like domains and chemotaxis systems

*Corresponding author

Email: bourret@med.unc.edu

Co-author email addresses:

Luke R. Vass: lukev3@vt.edu

Clay A. Foster: clay-foster@ouhsc.edu

Data Availability Statement:

The data that support the findings of this study were derived from the following resources available in the public domain: Pfam (release v.33), http://pfam.xfam.org; MiST, https://mistdb.com. The data we compiled are available as Datasets S1 through S4 in the Supplementary Material of this article.

Conflict of Interest Statement:

The authors declare no conflict of interest.

ABSTRACT

The ability to control locomotion in a dynamic environment provides a competitive advantage for microorganisms, thus driving the evolution of sophisticated regulatory systems. Nineteen known categories of chemotaxis systems control motility mediated by flagella and Type IV pili, plus other cellular functions. A key feature that distinguishes chemotaxis systems from generic two-component regulatory systems is separation of receptor and kinase functions into distinct proteins, linked by CheW scaffold proteins. This arrangement allows for formation of varied arrays with remarkable signaling properties. We recently analyzed sequences of CheW-like domains found in CheA kinases and CheW and CheV scaffold proteins. Sixteen *Architectures* of CheA, CheW, and CheV proteins contain ~94% of all CheW-like domains, forming six *Classes* with likely functional specializations.

We surveyed chemotaxis system categories and proteins containing CheW-like domains in ~1900 prokaryotic species, the most comprehensive analysis to date. The larger sample size revealed previously unknown insights. Co-occurrence analyses suggested that chemotaxis systems occur in non-random combinations within species, increasing our understanding of evolution of chemotaxis. Furthermore, many *Types* of CheW-like domains occurred predominantly with specific categories of chemotaxis systems, suggesting specialized functional interactions. For example, *Class* 2 (*Type* CheW.IC) domains exhibit properties spanning the primary *Classes* of CheW-like domains in CheA and CheW proteins. CheW.IC frequently co-occurred with methyl-accepting coiled coil (MAC) proteins, which contain both receptor and kinase functions. Although MAC proteins should not need CheW scaffolds to connect receptor and kinase functions, co-occurrence suggested that MAC systems may nevertheless benefit from array formation facilitated by CheW.IC domains.

KEYWORDS

Bacterial chemotaxis systems, CheW-like domains, CheA, CheW, CheV

1 | INTRODUCTION

Two component signaling systems are found in bacteria, archaea, and certain eukaryotes such as plants and fungi.[1,2] Two-component pathways allow organisms to sense and respond to environmental stimuli in an organized and timely manner (reviewed in [2]). The most basic two-component system consists of a membrane-bound sensor histidine kinase that binds ATP and modulates an autophosphorylation reaction in response to an external stimulus. The resulting phosphoryl group is transferred to an aspartate residue in the receiver domain of a downstream response regulator to elicit an appropriate cellular response. However, two-component pathways often exhibit more complexity, incorporating additional proteins to form branching signaling networks (reviewed in [3]). This increased complexity provides an opportunity to fine-tune the signal-response characteristics of a given pathway, tailoring the system to better suit the needs of the organism (various advantages are summarized in [4]).

The chemotaxis pathway is one of the most well-studied two-component systems[5] and is present in some form in nearly every motile microorganism. Chemotaxis is a regulatory strategy used to direct the movement of an organism towards resources (attractants) or away from undesirable substances (repellants). Variations on the chemotaxis system allow for locomotion in response to a variety of physicochemical parameters such as temperature, pH, magnetism, etc., in addition to nutrients.[6-10] The pathway utilizes a diverse repertoire of transmembrane environmental sensors to detect properties of interest. The sensors, known as chemoreceptors (also called methyl-accepting chemotaxis proteins, or MCPs), typically form mixed transmembrane arrays with remarkable, highly customizable signaling properties, including wide dynamic ranges, integration of mixed inputs, cooperativity, and rapid signal amplification potential.[11] The sophisticated information processing capabilities of the chemotaxis system are advantageous not only for general survival, but also for invasion-, colonization-, and virulence-related processes in pathogenic microorganisms.[12-16]

The chemotaxis pathway of *Escherichia coli* has been thoroughly characterized and is an example of a two-component system that incorporates additional proteins to achieve a more rapid and coordinated response.[17] The pathway begins at a transmembrane array of chemoreceptors. Activation of the sensor array depends on detection of an environmental stimulus and the methylation status of the receptors. Following detection, a stimulus is converted into receptor conformational changes to initiate processing and propagation. The signal is passed to the histidine kinase, CheA, which integrates information from multiple receptors through autophosphorylation after summing positive and negative stimuli.[18,19] The signal path then splits into “excitation” and “adaptation” branches. In the excitation path, phosphoryl groups are passed to the response regulator CheY. Phosphorylation alters the equilibria between active and inactive conformations in the CheY population, which ultimately modulates flagellar motor behavior and motility. The adaptation path in *E. coli* features CheR and CheB. CheR includes a methyltransferase domain that steadily adds methyl groups to the chemoreceptors, independent of environmental stimuli. CheB is a response regulator that includes a methylesterase domain whose activity is tightly regulated by phosphorylation and removes methyl groups from the chemoreceptors in response to a sufficient environmental change. The adaptation path forms a delayed negative feedback loop, imparting a “memory” to the system and allowing the organism to both follow a stimulus gradient and to reset upon reaching a uniform environment. Some chemotaxis systems also incorporate separate phosphatases, such as CheZ, to catalyze the removal of phosphoryl groups and terminate the response at a specific point in the pathway.[20] Many organisms encode multiple chemotaxis systems for regulating multiple forms of propulsion and/or gene expression[5].

An important distinction between a generic two-component pathway and a chemotaxis system is the separation of sensor and kinase functions into distinct protein species. Physical separation allows CheA kinases to integrate information from many different chemoreceptors, substantially enhancing the utility of the system.[18,19,21] This integration is facilitated by CheW proteins, which act as scaffolds between the various receptors and CheA kinases.[22,23] The classical architecture of CheA includes an histidine phosphotransfer (Hpt) domain (Pfam ID PF01627) containing the site of phosphorylation (a His residue), a dimerization domain (PF02895), an HATPase_c ATP binding and catalytic domain (PF02518), and a CheW-like domain (PF01584).[24] The CheW-like domain in *E. coli* CheA interacts with its counterpart CheW-like domain in standalone CheW proteins and also with cytoplasmic portions of the receptors to form MCP signaling arrays. The formation of supramolecular MCP•CheW•CheA oligomers is an essential part of the system and leads to CheA activation, signal propagation, and ultimately a downstream shift in flagellar behavior and/or locomotion.[21,23,25]

Most existing information on CheW-like domains describes the canonical, standalone CheW protein (the most abundant form in nature).[25] The structure of CheW consists of two connected β-barrels that form a bridge between the cytoplasmic regions of a receptor and the kinase CheA. The interactions between the free species of CheW and its partners (MCPs and CheAs) have been well-characterized for several microbial species, particularly *E. coli.*[26-30] However, the CheW-like domain itself is ubiquitous and can be found embedded in dozens of distinct architectures encompassing nearly every component and layer of the microbial chemotaxis system, e.g. fused to CheR or even CheZ.[31] CheW-like domains are evolutionarily related, regardless of their flanking architectures, and are relatively identifiable with traditional domain detection methods. However, the defining characteristics that distinguish the functionality of the domain within these different contexts are unknown.

The most common occurrence of CheW-like domains other than CheA- or CheW-lineage proteins (-lineage referring to proteins that can be loosely characterized as analogous to the canonical CheA and standalone CheW proteins in *E. coli*) is fused to a receiver domain in CheV proteins (reviewed in [22,32]). Various commonly studied organisms, including *Bacillus subtilis*, *Helicobacter pylori,* and *Vibrio cholerae,* encode one or more CheV proteins. While CheV proteins are present in approximately one-third of all chemotaxis systems, their role(s) are still poorly understood.[31] CheV is thought to be involved in both CheA modulation/MCP adaptation and array formation/polar localization.[33,34] The receiver domain of CheV may also serve as a general phosphate sink for the system.[32,35]

A landmark study by Wuichet and Zhulin established an evolutionary classification of chemotaxis signaling systems in prokaryotes[31], summarized here. The core components of essentially all chemotaxis systems, as outlined above, are the MCP•CheW•CheA arrays, the CheB and CheR adaptation enzymes, and CheY response regulators. Chemotaxis systems often have multiple MCPs and CheW proteins, and it is technically difficult to distinguish CheY from other single domain response regulators. Therefore, to provide a consistent foundation, the original classification was based on phylogenetic trees of CheA/CheB/CheR sequences and supported by certain other phylogenetic markers. There are 19 known categories of standard chemotaxis systems, each containing characteristic arrangements of *che* genes, distinct sets of auxiliary components (CheC phosphatase, CheD deamidase, CheV scaffold protein, CheX phosphatase, and/or CheZ phosphatase), and often unique architectures for certain core components. Seventeen categories of chemotaxis systems (F1 through F17) control flagellar motility, one controls Type IV pili (Tfp), and one controls alternative (non-motility) cellular functions (ACF). In addition, there are two categories of related chemotaxis systems based on methyl-accepting coiled coil (MAC) proteins, which contain both receptor and kinase functions, rather than separate MCP and CheA proteins.

A recent companion paper describes our classification (Figure 1) of the numerous architectural contexts of CheW-like domains and the implications thereof.[36] Nearly all (~94%) CheW-like domains are encompassed by 16 distinct *Architectures* (Figure 2). Because certain *Architectures* include multiple CheW-like domains, there are 21 major *Contexts* for CheW-like domains. The CheW.I *Architecture/Context* consists of three *Types* of sequences, whereas the other 20 *Contexts* each correspond to a single *Type*. As determined by the terminal level of our analysis, all 23 *Types* of CheW-like domain sequences sort into five or six *Classes*, likely related to specific functional specializations. Most CheW-like domains in CheW- and CheV-lineage proteins belong to *Class* 1 (*Type* CheW.IB). Most CheW-like domains in CheA-lineage proteins belong to *Class* 3, except for the CheA.VII and CheA.X *Architectures* (*Class* 4, which contain multiple Hpt domains) and the CheA.V.2 and CheA.VI.2 *Contexts* (*Class* 5, the C-terminal CheW-like domains in CheA proteins with two such domains). The rare (~1%) CheW.IC *Type* of CheW-like domains (*Class* 2) is found in CheA-lineage and CheW-lineage proteins and exhibits properties of both. About 20% of CheW-like domains in CheW-lineage proteins (*Class* 6, *Type* CheW.IA) appear subtly different from *Class* 1 and may form yet another specialized *Class*.

In this work, we combined our classification scheme of CheA-, CheW-, and CheV-lineage (CheW-like domain containing) proteins[36] with that of Wuichet and Zhulin for other chemotaxis proteins[31] to gain additional insights into the evolution and organization of chemotaxis systems. We found that (i) chemotaxis system categories occur in non-random combinations within microbial species, and (ii) specific *Architectures* of CheA/CheW proteins are preferentially associated with specific chemotaxis system categories, suggesting functional interactions.

2 | MATERIALS AND METHODS

2.1 | Protein sequence database and co-occurrence analysis

Analysis of CheW-like domains (PF01584) was described in [36], using sequences sourced from Representative Proteome 35 (RP35)[37] and the Pfam database (version 33, obtained May 2020).[38]

To analyze the co-occurrence (presence-absence) patterns of the various chemotaxis components encoded by proteomes within the RP35 sequence set, proteins containing one or more of the following domains were extracted: MCP (PF00015); CheR (PF01739); CheB (PF01339); CheD (PF03975); CheZ (PF04344); CheCX (also simply called CheC, PF04509/PF13690).[38] Receiver domain-containing proteins orthologous to CheY were excluded for the sake of interpretability. Full protein sequences (excluding the previously analyzed CheW-containing proteins) were scanned and classified with HMMER3 (version 3.3) using the previously described chemotaxis system models (utilized by the MiST database; version 3.0).[31,39-41] A combined frequency table was generated by merging the previously classified CheW-containing architectures with the other chemotaxis components by organism. Components with fewer than 20 positive occurrences were discarded from the analysis. The resulting count matrix of distinct chemotaxis proteins (encoded by 1887 distinct proteomes) was used to generate a heatmap (using the R package ComplexHeatmap, version 2.7.11) of the co-occurrence patterns within the representative proteomes.[42] The taxize package in R (version 0.9.99.947)[43] was used to assign putative phyla and classes to the batch of relevant organisms (sourced from the NCBI Taxonomy Browser[44]). Assignments were visualized as row/column annotations with ComplexHeatmap. Rows (components) and columns (species) were grouped in an unsupervised manner by hierarchical clustering, based on Spearman correlation, using Ward's clustering method (option “ward.D2”). The resulting dendrograms (both row and column) were split by height (using the cutree() function, as implemented by the row_split/column_split options) into putative functional “blocs” of chemotaxis components. Various heights were examined to optimize interpretability (data not shown).

To analyze the co-occurrence patterns of the chemotaxis classes themselves, rather than the individual components (see Figure 4; matching assignments made with the functional blocs in Figure 3), the original matrix generated to create Figure 3 was binarized. Rows featuring CheA, CheW, CheV and MCP paralogs were first removed. A binary scheme was then applied for each individual chemotaxis category (F1, F2, F4, F5, F6, F7, F7.z, F8, F9, F10, Tfp, ACF, MAC1, MAC2 and Uncat). If a given organism contained at least one of the corresponding components (CheB/C/D/R/Z) for a chemotaxis category, the category was considered present in the final table (=1). Those lacking were assigned absent (=0). A small percentage of organisms (<5%) lacked auxiliary components entirely and were excluded. A new heatmap was generated in a similar manner using ComplexHeatmap to re-cluster the proteomes (i.e., columns; 1797 distinct organisms). Row order was maintained to match Figure 3.

2.2 | Phyletic direct coupling analysis (PhyDCA) and network generation

The frequencies used to create the co-occurrence heatmap (covering the full, non-filtered complement of chemotaxis components) were also converted into a binary phylogenetic profile matrix to analyze pairwise evolutionary couplings. Data were analyzed with PhyDCA (using the mfDCA implementation) to estimate relevant phyletic pairings using a global statistical modelling approach.[45] The phyletic coupling (J_{ij}) between two domains and/or components in our data was used to estimate the favorability of finding multiple elements within the same species, corresponding to the principle that a biological process (i.e., chemotaxis) would require both components to function and produce a strong positive coupling. A negative coupling could also be interpreted as alternative solutions for similar functionality in a given system. We took the top 125 strongest positive pairwise couplings and created a non-directed graph for visualization purposes using the R packages igraph and ggraph (using the Fruchterman-Reingold algorithm).[46-48]

3 | RESULTS AND DISCUSSION

3.1 | Combinations of chemotaxis system categories are non-randomly distributed across prokaryotic species

We previously extracted all proteins that contained CheW-like domains and belonged to the 16 *Architectures* (Figure 2) that account for ~94% of CheW-like domains within the Representative Proteome 35 (RP35) dataset.[36] Using the same dataset, we extracted the remaining known chemotaxis proteins (with the exception of CheY), resulting in components from 1887 distinct proteomes. MCPs were classified by number of heptad repeats,[49] whereas CheB, CheC (including closely related CheX[50]), CheD, CheR, and CheZ proteins were assigned to the chemotaxis system categories of Wuichet and Zhulin.[31] We counted the number of components in each proteome and organized the data into a frequency matrix (Dataset S1), with organisms in columns and distinct chemotaxis components in rows. Hierarchical clustering was performed to optimally group organisms and chemotaxis proteins with similar co-occurrence patterns. The results were visualized in a composite heatmap (Figure 3).

The information presented by Figure 3 is challengingly dense, but salient features can be identified and discussed most easily using a grid coordinate system in which blocs are identified by assigned chemotaxis system class (rows, containing individual protein components; labeled on left with silver boxes) and representative proteome cluster (columns, containing distinct organisms; labeled on bottom with numbers as proteome clusters). First, Figure 3 does not include eight categories (F3 or F11 through F17) of chemotaxis systems. The aforementioned categories were rare in the surveyed proteomes (as previously observed[31]) and so were excluded from Figure 3 to facilitate interpretability. Second, the dominant feature of Figure 3 is that the data primarily clustered (in an unsupervised manner) into recognizable blocs, as a function of proteome and putative chemotaxis system category (row dendrogram not shown). Such a phenomenon strongly suggested that the data were linked in both dimensions (across chemotaxis system categories and across proteomes). We explore multiple aspects of the relationships between chemotaxis proteins (or components), chemotaxis system categories, and proteomes in the following sections.

If each species (proteome) encoded a single category of chemotaxis system, then two-dimensional clustering would be trivial, and proteomes would group perfectly into functional blocs by system category. However, >50% of all prokaryotic genomes that encode chemotaxis systems contain multiple systems (first determined in [31] and again corroborated by our work). With many known categories of chemotaxis systems, if combinations featuring multiple categories in a single species were random, then clustering by species would be disrupted, precluding the previously described functional blocs. The observed data structure in Figure 3 suggested that a restricted subset of common combinations of chemotaxis systems is evolutionarily favored. Preferred combinations must be either ancient (passed on to descendants of common ancestors) and/or are synergistically beneficial (arose independently multiple times). In either case, horizontal gene transfer of chemotaxis systems between species has not erased the pattern of combinatorial preferences in nature. Additionally, multiple chemotaxis systems present within the same species may serve as substrates for continuing evolution. For example, the modern class F7 chemotaxis system of *E. coli* evolved from a merger of the more ancient versions of class F6 and F7 systems.[51]

3.2 | Presence-absence analysis reveals evolutionarily favorable category combinations in organisms with multiple chemotaxis systems

By converting the frequency table of components used to generate Figure 3 into a simplified, binary presence/absence matrix featuring only the chemotaxis categories themselves, we next determined the most common naturally occurring combinations of chemotaxis system categories. Dataset S2 provides a full breakdown with corresponding proteome counts. The simplified matrix was also used to generate a heatmap displaying the presence/absence of each chemotaxis system category across species (Figure 4). We used Dataset S2 to calculate the total number of organisms encoding any given chemotaxis system category, irrespective of the presence/absence of other chemotaxis system categories (or paralogous instances of the same category), as well as the relative abundances of the various categories (Dataset S3).

We first sought to compare the distribution of chemotaxis system categories across species with the evolutionary relationships between the chemotaxis systems. The phylogenetic tree of chemotaxis systems features three main branches (here arbitrarily designated Branches 1, 2 and 3), with Branch 2 exhibiting three sub-Branches.[31] The most common 10% of chemotaxis category combinations observed in Dataset S2 accounted for two-thirds of the proteomes in our study and are displayed in relation to the various Branches in Table 1, with cross-referencing to their locations in Figure 4. Approximately a third of the proteomes encoded only a single category of chemotaxis system (Table 1, top).

We next focused on the proteomes encoding multiple chemotaxis systems, first examining pairwise combinations of categories found within the same Branch (Table 1, middle). In general, flagellar chemotaxis system categories within the same Branch (F1/F2, F4/F9/F10, and F7.z/F7/F8) co-occurred at substantially higher frequencies in Dataset S2 than would be expected based on frequencies of the constituent categories from Dataset S3 (calculations not shown). In contrast, the combination of categories F5/F6 occurred approximately five times less frequently than expected using the same relationship, implying some form of negative selection. It seems plausible that the components of the two categories (F5/F6) may interfere with one another. In contrast to the flagellar systems, categories ACF/Tfp and MAC1/2 both co-occurred at frequencies consistent with a random distribution. Overall, our findings imply that flagellar chemotaxis systems are not independent from one another, as might be expected if they control the same flagellar motors (e.g., having closely related CheY proteins to control the same motor could be advantageous). In contrast, ACF and Tfp systems appeared to act independently from each other, as did MAC1 and MAC2 systems.

Over 50% of all proteomes encoding multiple distinct chemotaxis systems in our dataset included categories from disparate Branches of the classification tree (Table 1 bottom and Dataset S2), suggesting highly diverse origins for most systems. We found that most outgroup combinations occurred at frequencies relatively consistent with a random distribution. A notable exception was class F7.z, which occurred with categories in outgroup Branches much less frequently than expected (consistent with a strong/semi-exclusive linkage between F7.z/F7 and F7.z/F8). Additionally, class F9 co-occurred with both F5 and F8 systems more frequently than expected, though we are uncertain as to the significance of this observation.

Finally, we examined the co-occurrent relationships between the flagellar chemotaxis system classes and the ACF/Tfp/MAC1/MAC2 systems. *A priori*, we speculated that the non-flagellar systems would operate independently from the flagellar-controlling classes, revealing no obvious selective pressure(s). While the majority of pairwise combinations between flagellar and non-flagellar systems (approximately 67%) supported our prediction, a full third deviated substantially (calculations not shown). One-quarter of pairwise combinations were observed at lower-than-expected frequencies, with nearly half of the cases of negative selection involving either F1 or F2 systems. Ten percent of pairwise combinations were observed at higher-than-expected frequencies, with nearly half of the cases of positive selection involving F10 systems. Once again, the significance of the deviating combinations is not immediately apparent.

The observations summarized in this section (and in Table 1/Figure 4) were only possible because we sampled proteomes from a large number (1887) of distinct organisms. Though members of some prokaryotic Phyla tend to encode particular categories of chemotaxis systems, the topologies of the chemotaxis and species classification trees do not match[31,52], implying different evolutionary paths. The primary combinations of chemotaxis system categories (Table 1) and their non-random nature (Figure 4) may provide clues into the evolution of chemotaxis systems.

3.3 | Matching *Architectures* of proteins containing CheW-like domains to preferred chemotaxis system categories

Examining the functional blocs revealed by clustering in Figure 3 suggested that various *Architectures* of CheA and CheW proteins are differentially favored by divergent chemotaxis systems. In particular, many *Architectures* clustered with components belonging to one chemotaxis system category. Wuichet and Zhulin described eight cases of distinct architectures for proteins containing CheW-like domains that were characteristic of specific chemotaxis system categories.[31] The four *Architectures* noted by Wuichet and Zhulin that were sufficiently abundant to be included in our study were CheA.III/CheA.IV, CheA.VI, CheA.XII, and CheW.III, which were linked to categories ACF/F3, F5, F4, and F9, respectively. Our data confirmed most previous assignments. Our larger sample size enabled us to also propose multiple additional assignments. Qualitative assignments of specific *Architectures* to specific chemotaxis systems inferred from sorting patterns in Figure 3 are summarized in Table 2. Quantitative analyses described in Section 3.4 strengthen and extend these observations. The new relationships identified in our work link *Architectures* CheA.I, CheA.II, CheA.V, CheA.VII, CheA.VIII, CheV.I, CheW.IA, CheW.IB, CheW.IC, and CheW.II with chemotaxis system categories F7/F8, F1, F5, F7.z, F7.z, F1/F6, F5/F6/F7.z, F1/F7/F8, MAC1, and F8/ACF, respectively. The additional assignments substantially expand our understanding of chemotaxis system organization.

Several individual rows in Figure 3 exhibited distributions from which we could glean additional insights. Prominent components spanning numerous proteome clusters included mcp.44H, mcp.24H, mcp.40H, mcp.34H, and mcp.36H (found in F1, F6, N/A, F8, and F7 blocs respectively); CheW.IB (F8 blocs); CheW.IA, which includes most CheW proteins (N/A blocs above F8); and CheA.I, the simplest and most common CheA *Architecture* (F7 blocs). The listed components are known to constitute the core signaling pathway shared by all chemotaxis systems. CheV.I (the sole version of CheV detected in any significant abundance; see [36]), found in nearly a third of all chemotaxis systems in nature, also spanned many proteome clusters (F6 blocs). Strikingly, the occurrence of CheV.I correlated well with the mcp.40H type of chemoreceptor, strongly suggesting preferential interaction(s) (N/A blocs above F8).

Figure 3 also revealed several key features shared by the MAC1/2 chemotaxis categories. Methyl-accepting coiled-coil (MAC) proteins are closely related to chemotaxis proteins but, to the best of our knowledge, have not been experimentally characterized in any respect. MAC proteins include apparent chemoreceptor and kinase domains, and either incorporate (MAC1) or are associated with (MAC2) CheB and CheR related domains.[31] It is unclear whether MAC proteins are evolutionary precursors of canonical chemotaxis systems or degenerate remnants. Proteome clusters 16, 17, and 18 featured high concentrations of organisms encoding MAC1 and/or MAC2 components but seemingly lacked other types of chemotaxis systems (MAC1 and MAC2 blocs). MAC systems were also scattered across numerous proteome clusters in Figure 3, rather than remaining constrained to contiguous blocs. Such a distribution suggested substantial phylogenetic prolificacy. Wuichet and Zhulin found that ~80% of species with MAC systems encode additional chemotaxis systems.[31] The distribution of MAC systems seen in Figure 3, based on our much larger sample size, supported and strengthened the original observation.

3.4 | Co-evolutionary analysis reveals functional communities of chemotaxis components

Due to the complex nature of the information represented in Figure 3 (and the relative inability of the human eye to untangle multivariate correlations), we sought a way to simplify the co-occurrence probabilities of individual components observed in the various chemotaxis systems in nature. The traditional approach to phylogenetic profiling utilizes some form of correlation metric, such as Hamming distance or Pearson correlation, to transform a simple binary matrix representing presence (1) or absence (0) of specific components/proteins/genes in various species into a corresponding interaction network. However, classical profiling suffers from several disadvantages, such as the influence of "intermediate" effects on apparent direct couplings (meaning that if A co-evolves with B, and B co-evolves with C, A may also appear to co-evolve with C). A more recent approach introduced the concept of direct coupling analysis, a statistical modeling technique able to distinguish between direct and more indirect co-evolutionary signals, to the profiling of presence-absence patterns.[45] This method, called Phyletic Direct Coupling Analysis, or PhyDCA, has demonstrated substantially increased accuracy compared to the more traditional correlation-based approaches and provided a convenient means by which to quantify the relationships presented in Figure 3. We converted our co-occurrence frequency table underlying Figure 3 (featuring the full list of chemotaxis components) into a simple binary presence-absence matrix and used PhyDCA to generate quantitative, pairwise phyletic couplings between individual components/domains. We then took the top 125 (~4%) strongest positive couplings (i.e., the presence of one component favors the presence of the other) and generated a non-directed graph to visualize the web of co-evolutionary signals (Figure 5). The complete list of pairwise coupling strengths (>3000 pairs) is in Dataset S4.

The relationships revealed by Figure 5 largely corroborated the clustering patterns of Figure 3 and the assignments made in Table 2. Individual chemotaxis components typically associated closely with others of the same system category (represented by shared colors in Figure 5), but less strongly to nodes outside the same group. The network representation was particularly useful for visualizing relationships between the disparate categories/chemoreceptors and the CheW-/CheV-lineage *Architectures*. Most of the chemotaxis categories could be traced to a least one CheW-lineage and CheA-lineage component in relatively short order. Category F2 components and mcp.48H appeared as a cluster unconnected to other components in Figure 5, but all exhibited couplings in the top ~10% to CheA.II (Dataset S4) and hence linked to the F1 chemotaxis system.

The strong connections between components observed in Figure 5 allowed us to confirm many of the chemotaxis system class assignments proposed in Table 2 (based on observations from Figure 3), as well as to infer several additional novel assignments. Key observations derived from Figure 5 are described in the following paragraphs.

Figure 3 shows four groups of components (labeled N/A or uncategorized) that sorted into isolated blocs rather than associating with the standard chemotaxis system categories. Figure 5 suggests that the components were not associated with one another through co-evolutionary processes, but rather were dispersed and associated with a diverse range of other proteins. One explanation for the differences between the results in Figures 3 and 5 is correlation of individual components with multiple chemotaxis system categories. Such a phenomenon would likely facilitate linkage in Figure 5 but confound the clustering procedure used for Figure 3.

CheW.IB (the most common *Type* of CheW) sorted with chemotaxis system category F8 in Figure 3 but connected primarily with the F1 and F7 systems in Figure 5.

This pattern is consistent with the fact that F1, F7, and F8 are the most abundant chemotaxis system categories (Dataset S3, Ref. [31]). CheW.IB linked with CheA.I and CheA.II (the two most common CheA *Architectures*[36]) in Figure 5. CheW.IB also connected to multiple types of MCPs in Figure 5. All described linkages are consistent with common core components utilized by many different categories of chemotaxis systems.

The CheW.IA *Type* comprises *Class* 6 of CheW-like domains and makes up ~20% of CheW.I proteins.[36] CheW.IA is subtly distinguishable from CheW.IB and CheV.I *Types* of CheW-like domains (*Class* 1) by some (but not all) methods of sequence analysis.[36] In Figure 5, CheW.IA made direct connections to various chemoreceptors, but not to any CheA proteins (i.e., phyletic coupling of CheW.IA was stronger to MCPs than to CheA-lineage proteins). We speculate that the distinction between Class 1 and Class 6 CheW-like domains is that the latter exhibit greater specificity or preference for interactions with certain classes of MCPs (e.g., mcp.40H, mcp.38H). Note that CheW.IA sorted with mcp.40H in Figure 3 (N/A bloc above F8), but not with a specific chemotaxis system class. In a related observation, CheW.IA and CheW.IB account for nearly all single domain CheW proteins in nature. Both made strong direct connections to mcp.40H in Figure 5. CheW.IB also made direct connections with mcp.44H (strong) and mcp.36H (weaker), and CheW.IA made a direct connection (strong) with mcp.38H. Collectively, MCPs from 36H, 38H, 40H and 44H account for nearly 90% of all MCPs in nature (at least as encompassed by the RP35 dataset). When viewed in this way, the phyletic couplings between such prolific components makes sense. However, we again speculate that the “unique” interactions noted for the CheW.IA and IB *Types* are rooted in preferences for distinct chemoreceptors. Although both likely share a robust ability to interact with mcp.40H, CheW.IA may also be able to interact with mcp.38H, whereas CheW.IB may be able to interact with both mcp.44H and mcp.36H.

CheW.IA was directly linked to F5, F6, and F7.z chemotaxis system categories in Figure 5. CheW.IB was directly linked to F1 and F7 components in Figure 5 and sorted with F8 in Figure 3. As the F7.z category diverged from F7 and became associated with CheA.VII and CheA.VIII rather than CheA.I *Architectures* (Table 1), the CheW *Type* may have diverged in parallel from CheW.IB (F7) to CheW.1A (F7.z).

A recent report noted that the F7 system of *E. coli* likely evolved from merging ancient versions of the F6 and F7 systems.[51] Wuichet and Zhulin were unable to assign a characteristic chemoreceptor to the F6 system category.[31] Whereas the heatmap in Figure 3 implies that the appropriate MCP may be mcp.24H, the phyletic coupling network suggests that it could also be mcp.40H, via an indirect (but relatively strong) correlation with CheW.IA. Similarly, category F7 is strongly associated with mcp.36H (as previously reported[31]) but may also be linked to mcp.40H via CheW.IB.

CheW.IC (*Class* 2) makes up ~1% of CheW-like domains and shares characteristics of CheW-like domains found in both CheA and CheW proteins.[36] CheW.IC did not sort with any known chemotaxis system category in Figure 3 (N/A bloc below F4), perhaps affected by its low abundance in nature. The same rarity makes interpretation of the distribution of CheW.IC in Figure 3 challenging. However, CheW.IC was strongly linked with the MAC1 category in Figure 5 (corroborated upon close inspection of Figure 3). MAC systems incorporate both receptor and kinase functions into a single protein species, implying that they do not need CheW scaffold proteins to bridge the two elements. It is not known if any MAC proteins form arrays in conjunction with CheW proteins. In principle, arrays of MAC proteins could provide previously described advantages (e.g., sensitive signal detection, amplification, integration) of a canonical chemoreceptor array, but array properties might be constrained by the one-to-one relationship between intramolecular chemoreceptor and kinase functions in MAC proteins. Arrays could also facilitate adaptation, for example by allowing the CheB and/or CheR domains of MAC1 proteins to modify adjacent receptors, or the separate CheR proteins of MAC2 systems to localize to the array by molecular brachiation.[53]

CheW.II sorted with ACF systems in Figure 2 but made its strongest connection to category F8 systems in Figure 5. In contrast to most other types of CheW proteins, CheW.II did not make strong direct connections to individual MCPs in Figure 5, implying that CheW.II proteins may be promiscuous and interact with multiple different types of chemoreceptors. An alternative interpretation equally consistent with the data is that CheW.II proteins do not interact with MCPs at all but have an as yet to be determined function. We are not aware of any experimental investigation of CheW.II proteins. CheA.IX did not sort with a specific chemotaxis system category in Figure 3 (N/A bloc below F4), but connected to CheW.II in Figure 5, suggesting an association with the category F8 and ACF systems. Similarly, CheW.III (category F9), CheA.XI (unassigned in Figure 3, N/A bloc below F4), and CheA.XII (category F4) were strongly interconnected in Figure 5, suggesting CheA.XI may belong to both the class F4 and F9 chemotaxis systems. Belonging to multiple categories of chemotaxis systems could explain a failure to sort coherently in Figure 3. There also may be functional reasons for these apparent interconnections. The CheA.IX *Architecture* lacks Hpt domains with phosphorylation sites, whereas the CheA.XI and CheA.XII *Architectures* lack the catalytic and ATP-binding HATPase_c domain (Figure 2). Therefore, all three must interact with other CheA proteins to participate in phosphotransfer reactions.

CheV.I clustered with category F6 chemotaxis systems in Figure 3 but also appeared coincident with the F1 category. Figure 5 confirmed connections between CheV.I and the F1 and F6 categories. Figure 3 also showed a strong correlation between CheV.I and mcp.40H, which was again confirmed by Figure 5. However, CheV.I did not show a direct connection to any specific CheA *Architecture*. Curiously, besides a link with mcp.40H, the only other strong direct correlations formed by CheV.I involved the phosphatases chez.F6/chec.F1 and the methyltransferase cher.F1. The role(s) of CheV-lineage proteins and their attached receiver domains are poorly understood. Some evidence suggests that CheV is involved in the chemotaxis adaptation process[32], making the correlation between CheV.I and the CheC/CheR components of the F1 class[31] understandable. However, the nature of the connection between CheV.I and chez.F6 is less clear and raises the concept of CheZ (and possibly the CheC of category F1) acting upon the attached phosphorylatable receiver domain of CheV.I in the capacity of a phosphatase. In fact, CheZ has phosphatase activity toward one of the three CheV proteins in *H. pylori*.[54] It is not known whether CheZ distinguishes between different CheV proteins based on their CheW-like and/or receiver domains.

The use of the PhyDCA approach has several disadvantages that must be considered. One is that the observed phyletic couplings do not necessarily correspond to direct biophysical interactions. Strong coupling may also represent events such as genomic co-localization (a limitation the original authors circumvented by including an additional residue-level covariance analysis to predict likely direct interaction partners).[45] Because of the narrow perspective of our study (i.e., focusing on chemotaxis systems, rather than entire proteomes), we believe that this disadvantage was minimized. However, as can be seen in Figure 3, paralogous chemotaxis components are common in bacteria, particularly for chemoreceptors. Introducing a residue-level analysis step may facilitate the untangling of specific paralog interactions,[45,55] especially matching the various *Types* of CheW-like domains in CheA-, CheW-, and CheV-lineage proteins to their partner MCP components. Such an analysis would provide additional insight into the diverse chemotaxis systems found in nature. Additionally, PhyDCA (and many other phylogenetic profiling methods) relies on a binary presence/absence data structure to simplify data processing and interpretation. Excluding the substantial amount of paralogous protein data found in our microbial dataset very likely ignores valuable co-occurrence information. However, our two-pronged approach to the problem (using the full co-occurrence matrix to identify functional blocs/clusters in Figure 3 and using the transformed binary profile to create the simplified network representation in Figure 5) likely mitigates the issue.

3.5 | Negative phyletic couplings reveal putative overlapping functionality among specific chemotaxis components

The PhyDCA model can also be used to predict negative phyletic couplings, i.e., the presence of one component in a proteome disfavors the presence of another.[45] Logic suggests that components in such a scenario likely share overlapping (or at least closely related) functionalities (sometimes referred to as “alternative” solutions). For example, the top negative coupling within Dataset S4 involved the components cher.F8 and cher.Uncat, presumably because both methyltransferases serve highly similar functions. We sought to use the negative pairings to identify the overlapping roles of the more unusual *Architectures* containing CheW-like domains. The third strongest negative coupling involved CheW.II and CheW.III, leading to several related observations. None of the top negative phyletic pairings involving CheW.II or CheW.III feature any other CheW-lineage *Architecture*, implying, along with frequent co-occurrences with other CheW proteins in Figure 3, that the highly unusual CheW.II/III *Architectures* are not “alternative” solutions for the more standard CheW-lineage components (i.e., CheW.II does not replace two distinct single-domain CheW proteins), but likely serve novel functionalities as a result of some form of convergent evolution. Curiously, few other instances of anticorrelated components with presumably similar functions are present in the list of top negative phyletic pairs, with most entries involving disparate component types (and are therefore not likely to be consequences of convergent evolution). The only exceptions (from the top 4% strongest anticorrelated phyletic pairs) were CheW.IA with CheW.IC (both CheW-lineage scaffolds), chec.F1 with chez.F7 (both phosphatases), CheA.I with CheA.II (the two most abundant CheA-lineage kinases), cher.F10 with cher.F5 (both methyltransferases), cheb.F1 with cheb.F10 (both methylesterases), cher.F1 with cher.F8 (both methyltransferases), cheb.F7 with cheb.F8 (both methylesterases), and finally cheb.F10 with cheb.F5 (both methylesterases).

3.6 | Insights into evolution and organization of prokaryotic chemotaxis systems

CheW-like domains play a central role in the signal transduction systems that regulate prokaryotic chemotaxis by linking receptors and kinases into large arrays. Almost all CheW-like domains occur in a limited number of *Architectures* of CheA-, CheW-, and CheV-lineage proteins (Figure 2).[36] Furthermore, CheW-like domains have evolved into distinct functional *Classes* (Figure 1).[36] We inventoried chemotaxis proteins encoded by ~1900 species (Dataset S1) and examined their distribution in two dimensions: by chemotaxis system category[31] and by species. Successful unsupervised clustering of components into blocs (Figure 3) strongly suggested that the components were linked in both dimensions, leading to two central conclusions. First, combinations of chemotaxis systems encoded by individual species tend to be non-random (Figure 4, Table1, Dataset S2). Specific co-occurrence patterns and frequencies (Dataset S3) should provide insights into evolution of chemotaxis systems. Second, we inferred probable functional associations between each *Architecture* of CheA-, CheW-, and CheV-lineage proteins and specific categories of chemotaxis systems (Figure 5, Table 2, Dataset S4). These assignments lay a foundation for future investigations into the mechanisms that underly apparent functional specialization of different chemotaxis protein *Architectures*.

ACKNOWLEDGEMENTS

We thank Emily N. Kennedy and Sarah A. Barr for their insightful input and helpful discussions.

This work was funded by National Institutes of Health grant GM050860 to Robert B. Bourret. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

REFERENCES

1. Alvarez AF, Barba-Ostria C, Silva-Jiménez H, Georgellis D. Organization and mode of action of two component system signaling circuits from the various kingdoms of life. *Environ Microbiol.* 2016;18(10):3210-3226.
2. Zschiedrich CP, Keidel V, Szurmant H. Molecular mechanisms of two-component signal transduction. *J Mol Biol.* 2016;428(19):3752-3775.
3. Gao R, Stock AM. Biological insights from structures of two-component proteins. *Annu Rev Microbiol.* 2009;63(1):133-154.
4. Bourret RB, Kennedy EN, Foster CA, Sepúlveda VE, Goldman WE. A radical reimagining of fungal two-component regulatory systems. *Trends Microbiol.* 2021;29(10):883-893.
5. Kirby JR. Chemotaxis-like regulatory systems: Unique roles in diverse bacteria. *Annu Rev Microbiol.* 2009;63(1):45-59.
6. Popp F, Armitage JP, Schüler D. Polarity of bacterial magnetotaxis is controlled by aerotaxis through a common sensory pathway. *Nat Commun.* 2014;5(1):5398.
7. Tohidifar P, Plutz MJ, Ordal GW, Rao CV. The mechanism of bidirectional pH taxis in *Bacillus subtilis*. *J Bacteriol.* 2020;202(4).
8. Vaknin A, Berg HC. Osmotic stress mechanically perturbs chemoreceptors in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 2006;103(3):592-596.
9. Yang Y, Sourjik V. Opposite responses by different chemoreceptors set a tunable preference point in *Escherichia coli* pH taxis. *Mol Microbiol.* 2012;86(6):1482-1489.
10. Yoney A, Salman H. Precision and variability in bacterial temperature sensing. *Biophy J.* 2015;108(10):2427-2436.

11. Parkinson JS, Hazelbauer GL, Falke JJ. Signaling and sensory adaptation in *Escherichia coli* chemoreceptors: 2015 update. *Trends Microbiol.* 2015;23(5):257-266.

12. Yao J, Allen C. Chemotaxis is required for virulence and competitive fitness of the bacterial wilt pathogen *Ralstonia solanacearum*. *J Bacteriol.* 2006;188(10):3697-3708.

13. Butler SM, Camilli A. Going against the grain: Chemotaxis and infection in *Vibrio cholerae*. *Nat Rev Microbiol.* 2005;3(8):611-620.

14. Chandrashekhar K, Kassem II, Rajashekara G. *Campylobacter jejuni* transducer like proteins: Chemotaxis and beyond. *Gut Microbes.* 2017;8(4):323-334.

15. Josenhans C, Suerbaum S. The role of motility as a virulence factor in bacteria. *Int J Med Microbiol.* 2002;291(8):605-614.

16. Kreling V, Falcone FH, Kehrenberg C, Hensel A. *Campylobacter sp.*: Pathogenicity factors and prevention methods—new molecular targets for innovative antivirulence drugs? *Appl Microbiol Biotechnol.* 2020;104(24):10409-10436.

17. Sourjik V, Wingreen NS. Responding to chemical gradients: Bacterial chemotaxis. *Curr Opin Cell Biol.* 2012;24(2):262-268.

18. Khan S, Spudich JL, McCray JA, Trentham DR. Chemotactic signal integration in bacteria. *Proc Natl Acad Sci U S A.* 1995;92(21):9757-9761.

19. Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA. The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. *Annu Rev Cell Dev Biol.* 1997;13:457-512.

20. Zhao R, Collins EJ, Bourret RB, Silversmith RE. Structure and catalytic mechanism of the *E. coli* chemotaxis phosphatase CheZ. *Nat Struct Biol.* 2002;9(8):570-575.

21. Bi S, Sourjik V. Stimulus sensing and signal processing in bacterial chemotaxis. *Curr Opin Microbiol.* 2018;45:22-29.

22. Huang Z, Pan X, Xu N, Guo M. Bacterial chemotaxis coupling protein: Structure, function and diversity. *Microbiol Res.* 2019;219:40-48.

23. Pinas GE, DeSantis MD, Cassidy CK, Parkinson JS. Hexameric rings of the scaffolding protein CheW enhance response sensitivity and cooperativity in *Escherichia coli* chemoreceptor arrays. *Sci Signal.* 2022;15(718):eabj1737.

24. Bilwes AM, Alex LA, Crane BR, Simon MI. Structure of CheA, a signal-transducing histidine kinase. *Cell.* 1999;96(1):131-141.

25. Park S-Y, Borbat PP, Gonzalez-Bonet G, et al. Reconstruction of the chemotaxis receptor–kinase assembly. *Nat Struct Mol Biol.* 2006;13(5):400-407.

26. Ortega DR, Mo G, Lee K, et al. Conformational coupling between receptor and kinase binding sites through a conserved salt bridge in a signaling complex scaffold protein. *PLOS Comp Biol.* 2013;9(11):e1003337.

27. Wang X, Vu A, Lee K, Dahlquist FW. CheA-receptor interaction sites in bacterial chemotaxis. *J Mol Biol.* 2012;422(2):282-290.

28. Boukhvalova MS, Dahlquist FW, Stewart RC. CheW binding interactions with CheA and Tar: Importance for chemotaxis signaling in *Escherichia coli*. *J Biol Chem.* 2002;277(25):22251-22259.

29. Bellenger K, Ma X, Shi W, Yang Z. A CheW homologue is required for *Myxococcus xanthus* fruiting body development, social gliding motility, and fibril biogenesis. *J Bacteriol.* 2002;184(20):5654-5660.

30. 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(6):1261-1272.

31. Wuichet K, Zhulin IB. Origins and diversification of a complex signal transduction system in prokaryotes. *Sci Signal.* 2010;3(128):ra50-ra50.

32. Alexander RP, Lowenthal AC, Harshey RM, Ottemann KM. CheV: CheW-like coupling proteins at the core of the chemotaxis signaling network. *Trends Microbiol.* 2010;18(11):494-503.

33. Lowenthal AC, Simon C, Fair AS, et al. A fixed-time diffusion analysis method determines that the three *cheV* genes of *Helicobacter pylori* differentially affect motility. *Microbiology.* 2009;155(Pt 4):1181-1191.

34. Yang W, Alvarado A, Glatter T, Ringgaard S, Briegel A. Baseplate variability of *Vibrio cholerae* chemoreceptor arrays. *Proc Natl Acad Sci U S A.* 2018;115(52):13365-13370.

35. Ortega DR, Zhulin IB. Evolutionary genomics suggests that CheV Is an additional adaptor for accommodating specific chemoreceptors within the chemotaxis signaling complex. *PLOS Comp Biol.* 2016;12(2):e1004723.

36. Vass LR, Bascum KM, Bourret RB, Foster CA. Generalized strategy to analyze domains in the context of parent protein architecture: Case study of CheW. *Proteins.* 2022;Submitted.

37. Chen C, Natale DA, Finn RD, et al. Representative proteomes: A stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One.* 2011;6(4):e18910-e18910.

38. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2020;49(D1):D412-D419.

39. Gumerov VM, Ortega DR, Adebali O, Ulrich LE, Zhulin IB. MiST 3.0: An updated microbial signal transduction database with an emphasis on chemosensory systems. *Nucleic Acids Res.* 2019;48(D1):D459-D464.

40. Wuichet K, Alexander RP, Zhulin IB. Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis. *Methods Enzymol.* 2007;422:1-31.

41. Eddy SR. Accelerated Profile HMM Searches. *PLOS Comp Biol.* 2011;7(10):e1002195.

42. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.* 2016;32(18):2847-2849.

43. Chamberlain SA, Szocs E. taxize: taxonomic search and retrieval in R. *F1000Res.* 2013;2:191.

44. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford).* 2020;2020.

45. Croce G, Gueudré T, Ruiz Cuevas MV, et al. A multi-scale coevolutionary approach to predict interactions between protein domains. *PLOS Comp Biol.* 2019;15(10):e1006891.

46. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. *Software Pract Exper.* 1991;21(11):1129-1164.

47. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network visualizations of relationships in psychometric data. *J Stat Softw.* 2012;48(4):1 - 18.

48. *ggraph: An implementation of grammar of graphics for graphs and networks* [computer program]. Version 2.0.52021.

49. Alexander RP, Zhulin IB. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci U S A.* 2007;104(8):2885-2890.

50. Silversmith RE. Auxiliary phosphatases in two-component signal transduction. *Curr Opin Microbiol.* 2010;13(2):177-183.

51. Ortega DR, Yang W, Subramanian P, et al. Repurposing a chemosensory macromolecular machine. *Nat Commun.* 2020;11(1):2041.

52. Gumerov VM, Andrianova EP, Zhulin IB. Diversity of bacterial chemosensory systems. *Curr Opin Microbiol.* 2021;61:42-50.

53. Levin MD, Shimizu TS, Bray D. Binding and diffusion of CheR molecules within a cluster of membrane receptors. *Biophys J.* 2002;82(4):1809-1817.

54. Lertsethtakarn P, Ottemann KM. A remote CheZ orthologue retains phosphatase function. *Mol Microbiol.* 2010;77(1):225-235.

55. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Natl Acad Sci U S A.* 2016;113(43):12186-12191.

56. Adapted from “Flow Chart (3 Levels, Vertical) 3”, by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

TABLE 1. Most common chemotaxis system category combinations in Representative Proteomes[a]

Br. 2A[b]		Br. 2C		Branch 2B			Branch 1			Branch 3						Category	% of	RP Cluster
F1	F2	F5	F6	F4	F9	F10	F7.z[c]	F7	F8	ACF	Tfp	MAC1	MAC2	Uncat	Count	Combination	Proteomes	in Figure 4
Single systems:																		
+															304	F1	16.9	6
		+													87	F5	4.8	12 right
			+												18	F6	1.0	10 middle
								+							74	F7	4.1	8 right
										+					29	ACF	1.6	15
											+				17	Tfp	1.0	9 left
												+			106	MAC1	5.9	18 left
													+		46	MAC2	2.6	14 right
														+	9	Uncat	0.5	3 right
Combinations within the same Branch:																		
+	+														9	F1 + F2	0.5	6 left
							+	+	+						44	F7.z + F7 + F8	2.5	7 middle
							+	+							32	F7 + F7.z	1.8	7 left
								+	+						14	F7 + F8	0.8	8 left
												+	+		46	MAC1 + MAC2	2.6	17 left
Combinations between Branches:																		
+								+							95	F1 + F7	5.3	4 right
+		+													28	F1 + F5	1.6	5 left
+			+												12	F1 + F6	0.7	10 left
+					+										19	F1 + F9	1.1	1
+												+			16	F1 + MAC1	0.9	18 right
+												+	+		14	F1 + MAC1 + MAC2	0.8	17 right
+													+		56	F1 + MAC2	3.1	14 left
+														+	17	F1 + Uncat	1.0	3 left
		+						+							19	F5 + F7	1.1	12 middle
		+										+			20	F5 + MAC1	1.1	12 left
			+					+							11	F6 + F7	0.6	10 right
			+					+			+				9	F6 + F7 + Tfp	0.5	9 right

			+	+	28	ACF + MAC1	1.6	16
+	+	+		+	14	F7.z + F7 + F8 + MAC1	0.8	7 right
+	+	+	+		9	F7.z + F7 + F8 + ACF	0.5	7 middle
	+			+	24	F7 + MAC1	1.3	8 middle
	+	+		+	16	F7 + F8 + MAC1	0.9	8 left

[a] From Dataset S2, which includes 1797 proteomes. The 90% of combinations that each represent < 0.5% of the total dataset are not shown in this table.

[b] The phylogenetic tree in Figure 7 of Wuichet & Zhulin[31] that forms the basis for classification of chemotaxis system categories has three main branches, arbitrarily numbered here. Branch 2 has three main sub-branches.

[c] The subset of F7 systems that contain CheZ. See Figure 3.

TABLE 2. Assignment of CheA and CheW protein *Architectures* to chemotaxis system categories

Protein	CheW-like	Chemotaxis	Evidence for Chemotaxis System Assignment		
Architecture[a]	Domain *Class*[b]	System Category[c]	Figure 6 of [31]	Figure 3	Figure 5
CheA.I	3	F7		+	+
CheA.I	3	F8			+
CheA.II	3	F1		+	+
CheA.III	3	ACF	+	+	+
CheA.III	3	F3	+[d]		
CheA.IV	3	ACF	+	+	+
CheA.IV	3	F3	+[d]		
CheA.V	3, 5	F5	[e]	+	+
CheA.VI	3, 5	F5	+	+	+
CheA.VII	4	F7.z[f]	[e]	+	+
CheA.VIII	3	F7.z	[e]	+	+
CheA.IX	3	Uncertain (F8 and ACF?)			+
CheA.X	4	Unassigned[g]	[e]		
CheA.XI	3	Uncertain (F4 and F9?)			+
CheA.XII	3	F4	+	+	+
CheV.I	1	F6		+	+
CheV.I	1	F1			+
CheW.IA	6	F5, F6, F7.z			+
CheW.IB	1	F8		+	
CheW.IB	1	F1, F7			+
CheW.IC	2	MAC1			+
CheW.II	1	ACF	[e]	+	
CheW.II	1	F8	[e]		+
CheW.III	1	F9	+	+	+

[a] Outlined in Figure 2. The CheW.I *Architecture* splits into three *Types*.[36]

[b] From [36]. The two CheW-like domains in CheA.V and CheA.VI *Architectures* belong to different *Classes*.

[c] Our sample does not contain enough representatives for analysis of system categories F3, F11-F17.

[d]Wuichet & Zhulin[31] noted CheA proteins modified only by C-terminal receiver domains (CheA.III or CheA.IV) were consistently observed in F3 chemotaxis systems. However, our PhyDCA scores in Dataset S4 do not support linkage of either the CheA.III or CheA.IV *Architectures* to either F3 or F4 chemotaxis systems.

[e]Although the CheA.V, CheA.VII, CheA.VIII, CheA.X, and CheW.II *Architectures* contain additional domains with respect to canonical CheA or CheW *Architectures* and are sufficiently abundant to be included in [36], these *Architectures* were apparently not observed sufficiently consistently in specific chemotaxis system categories to be noted by Wuichet & Zhulin[31], who analyzed a much smaller sample size.

[f]The subset of F7 chemotaxis systems that contain CheZ. See Figure 3.

[g]Did not sort with a specific chemotaxis system category in Figure 3 and did not make sufficiently strong connections to be included in Figure 5.

FIGURE LEGENDS

FIGURE 1. Summary of classification scheme for CheW-like domains used in [36]. Note that Wuichet & Zhulin refer to 19 different kinds of chemotaxis systems as “classes”.[31] To avoid confusion, in this report we use “*Classes*” for CheW-like domains and “categories” for chemotaxis systems. Created with BioRender.com.[56]

FIGURE 2. Major *Architectures* of proteins that contain CheW-like domains (from [36]). CheA-lineage *Architectures* are designated by a Roman numeral suffix in order of decreasing abundance. CheW-lineage *Architectures* are designated by a Roman numeral suffix indicating the number of CheW-like domains. For *Architectures* with multiple CheW-like domains, the *Contexts* of CheW-like domains within an *Architecture* are designated by an Arabic numeral suffix indicating N- to C-terminal order (not shown). The CheW.I *Context* includes sequences of three distinct *Types*, designated CheW.IA, CheW.IB, and CheW.IC (not shown).

FIGURE 3. Co-occurrences of individual chemotaxis components in RP35 representative proteome set. Total occurrence counts were used. Components with < 20 occurrences were excluded. Column annotations were shaded by Phyla (groups with < 10 occurrences are unlabeled) and Class (groups with < 10 occurrences are unlabeled). Notable organisms were tagged. Results were split by dendrogram height into functional “blocs” by clustering both proteomes (columns) and chemotaxis components (rows). Representative Proteome clusters were labeled as 1-18, whereas component blocs were labeled with most likely chemotaxis system category. Row dendogram not shown.

FIGURE 4. Simplified co-occurrence schematic of chemotaxis system categories in RP35 representative proteome set. A binary presence/absence scheme was used for visualization. A chemotaxis system category was determined to be present in a given proteome if at least one of the following components was detected of the appropriate category: CheB/C/D/R/Z. CheA/V/W and MCP components were excluded from the analysis, because some of these components function with more than one chemotaxis system category. Notable organisms were labelled. Row order was maintained for consistency with Figure 3. Results were split by dendrogram height into functional “blocs” by clustering proteomes (columns). However, because the datasets upon which Figures 3 and 4 are based are different, the resulting proteome clusters and cluster numbers are different than in Figure 3.

FIGURE 5. Network representation of inferred phyletic couplings between *Architectures* containing CheW-like domains and remaining chemotaxis system components. Co-occurrence data of chemotaxis components extracted from the RP35 representative proteome set were converted to a binary phylogenetic profile matrix. Phyletic Direct Coupling Analysis (PhyDCA) was used to quantify the favorability (correlation) of chemotaxis components co-occurring within the same organism. Strong favorability/high coupling typically corresponds to a cellular function (i.e., chemotaxis) requiring both components, though not necessarily to a direct biophysical interaction. The top 125 positive co-evolutionary pairings were used to construct a graph based on phyletic coupling strength. Architectural assignments correspond to those included in Figure 1 (i.e., identical thresholds). Edge width was scaled with phyletic coupling strength. Notes: *Architecture* CheA.X did not appear in the top 125 strongest phyletic

couplings and was excluded from the graph. cheb.F2, cher.F2, and mcp.48H formed a cluster disconnected from the rest of the network, but all three coupled to CheA.II at slightly lower strengths (top ~10%) (Dataset S4).

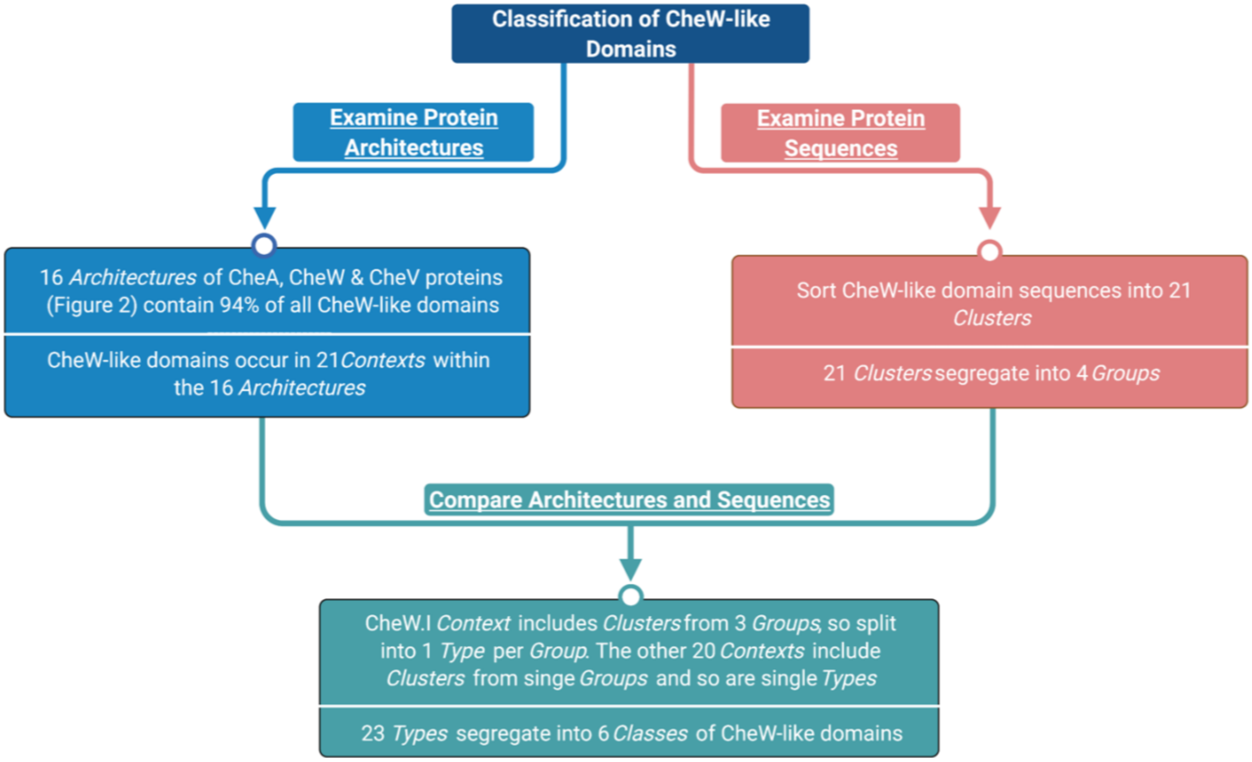

Classification of CheW-like Domains
Examine Protein Architectures
Examine Protein Sequences
16 *Architectures* of CheA, CheW & CheV proteins (Figure 2) contain 94% of all CheW-like domains
CheW-like domains occur in 21 *Contexts* within the 16 *Architectures*
Sort CheW-like domain sequences into 21 *Clusters*
21 *Clusters* segregate into 4 *Groups*
Compare Architectures and Sequences
CheW.I *Context* includes *Clusters* from 3 *Groups*, so split into 1 *Type* per *Group*. The other 20 *Contexts* include *Clusters* from singe *Groups* and so are single *Types*
23 *Types* segregate into 6 *Classes* of CheW-like domains

FIGURE 1. Summary of classification scheme for CheW-like domains used in [36]. Note that Wuichet & Zhulin refer to 19 different kinds of chemotaxis systems as "classes".[31] To avoid confusion, in this report we use "*Classes*" for CheW-like domains and "categories" for chemotaxis systems. Created with BioRender.com.[56]

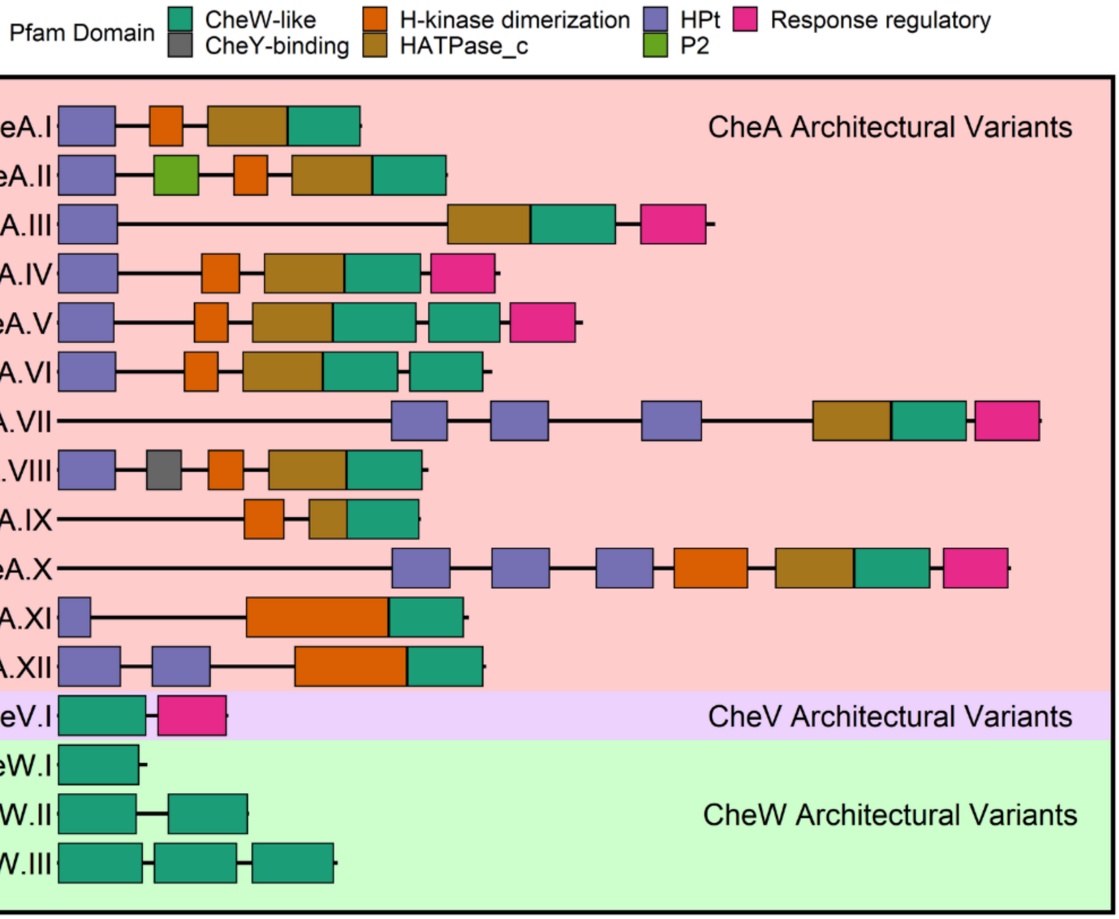


FIGURE 2. Major *Architectures* of proteins that contain CheW-like domains (from [36]). CheA-lineage *Architectures* are designated by a Roman numeral suffix in order of decreasing abundance. CheW-lineage *Architectures* are designated by a Roman numeral suffix indicating the number of CheW-like domains. For *Architectures* with multiple CheW-like domains, the *Contexts* of CheW-like domains within an *Architecture* are designated by an Arabic numeral suffix indicating N- to C-terminal order (not shown). The CheW.I *Context* includes sequences of three distinct *Types*, designated CheW.IA, CheW.IB, and CheW.IC (not shown).

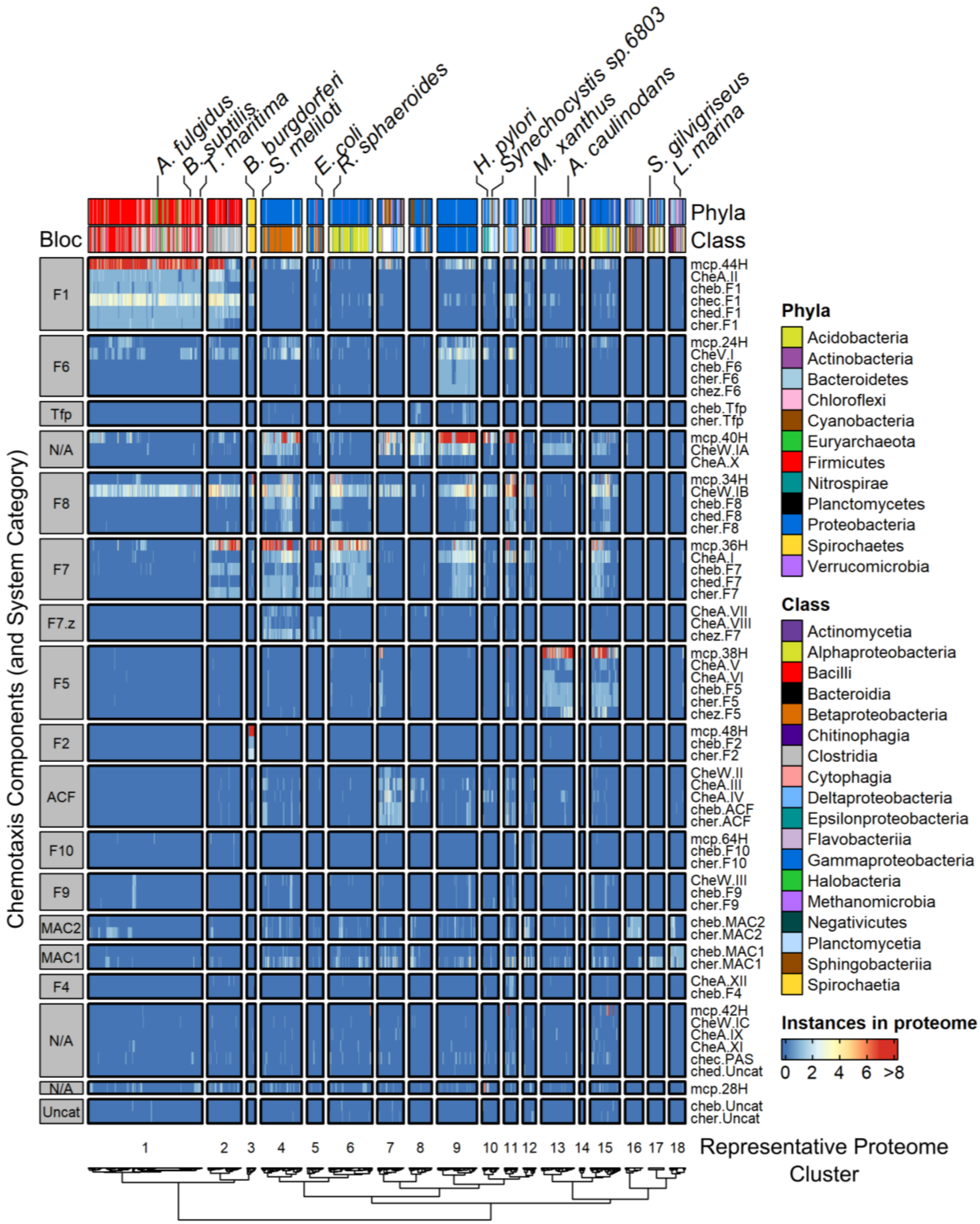
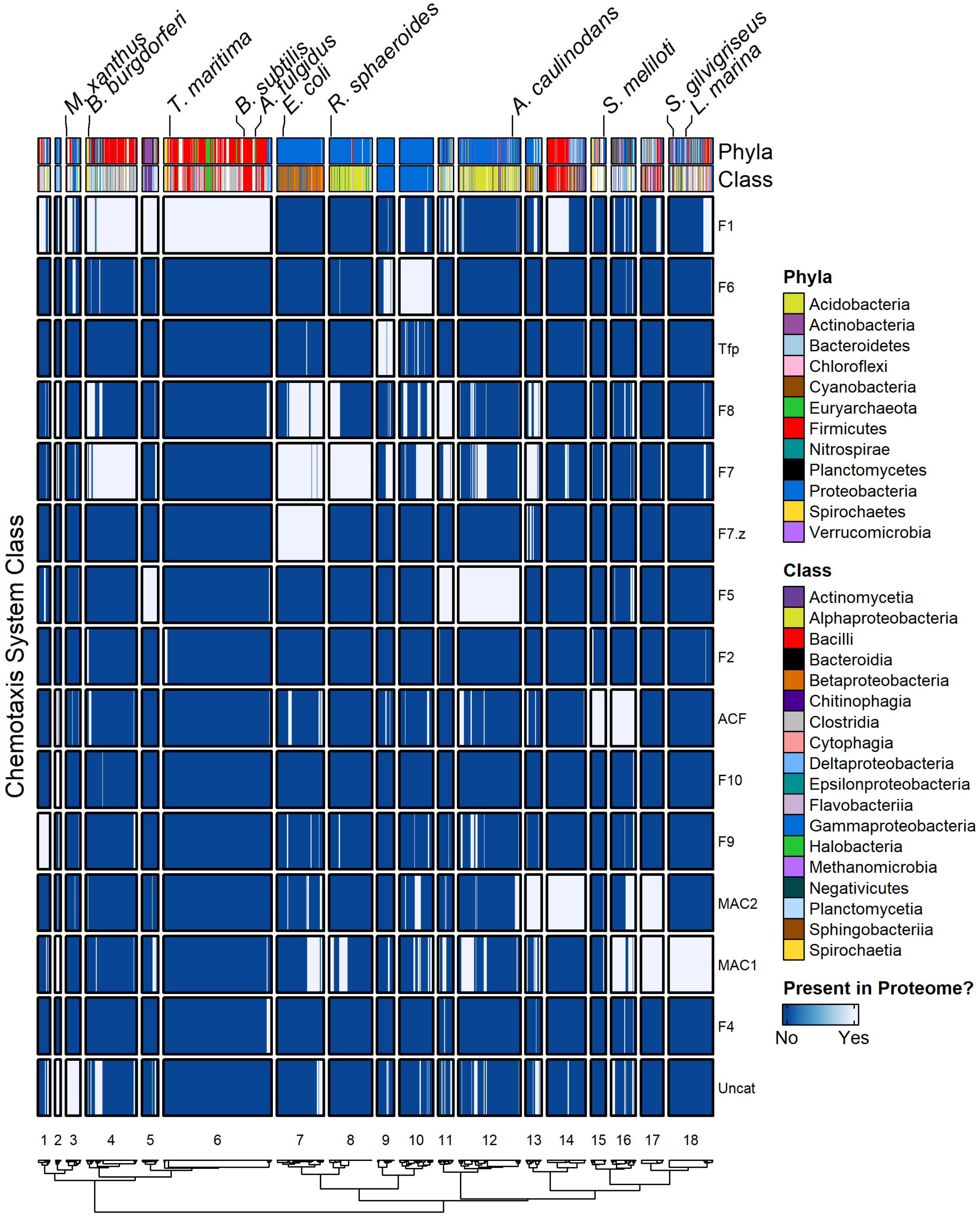


FIGURE 3. Co-occurrences of individual chemotaxis components in RP35 representative proteome set. Total occurrence counts were used. Components with <

20 occurrences were excluded. Column annotations were shaded by Phyla (groups with < 10 occurrences are unlabeled) and Class (groups with < 10 occurrences are unlabeled). Notable organisms were tagged. Results were split by dendrogram height into functional “blocs” by clustering both proteomes (columns) and chemotaxis components (rows). Representative Proteome clusters were labeled as 1-18, whereas component blocs were labeled with most likely chemotaxis system category. Row dendogram not shown.



908

909 **FIGURE 4. Simplified co-occurrence schematic of chemotaxis system categories**
910 **in RP35 representative proteome set.** A binary presence/absence scheme was used

for visualization. A chemotaxis system category was determined to be present in a given proteome if at least one of the following components was detected of the appropriate category: CheB/C/D/R/Z. CheA/V/W and MCP components were excluded from the analysis, because some of these components function with more than one chemotaxis system category. Notable organisms were labelled. Row order was maintained for consistency with Figure 3. Results were split by dendrogram height into functional “blocs” by clustering proteomes (columns). However, because the datasets upon which Figures 3 and 4 are based are different, the resulting proteome clusters and cluster numbers are different than in Figure 3.

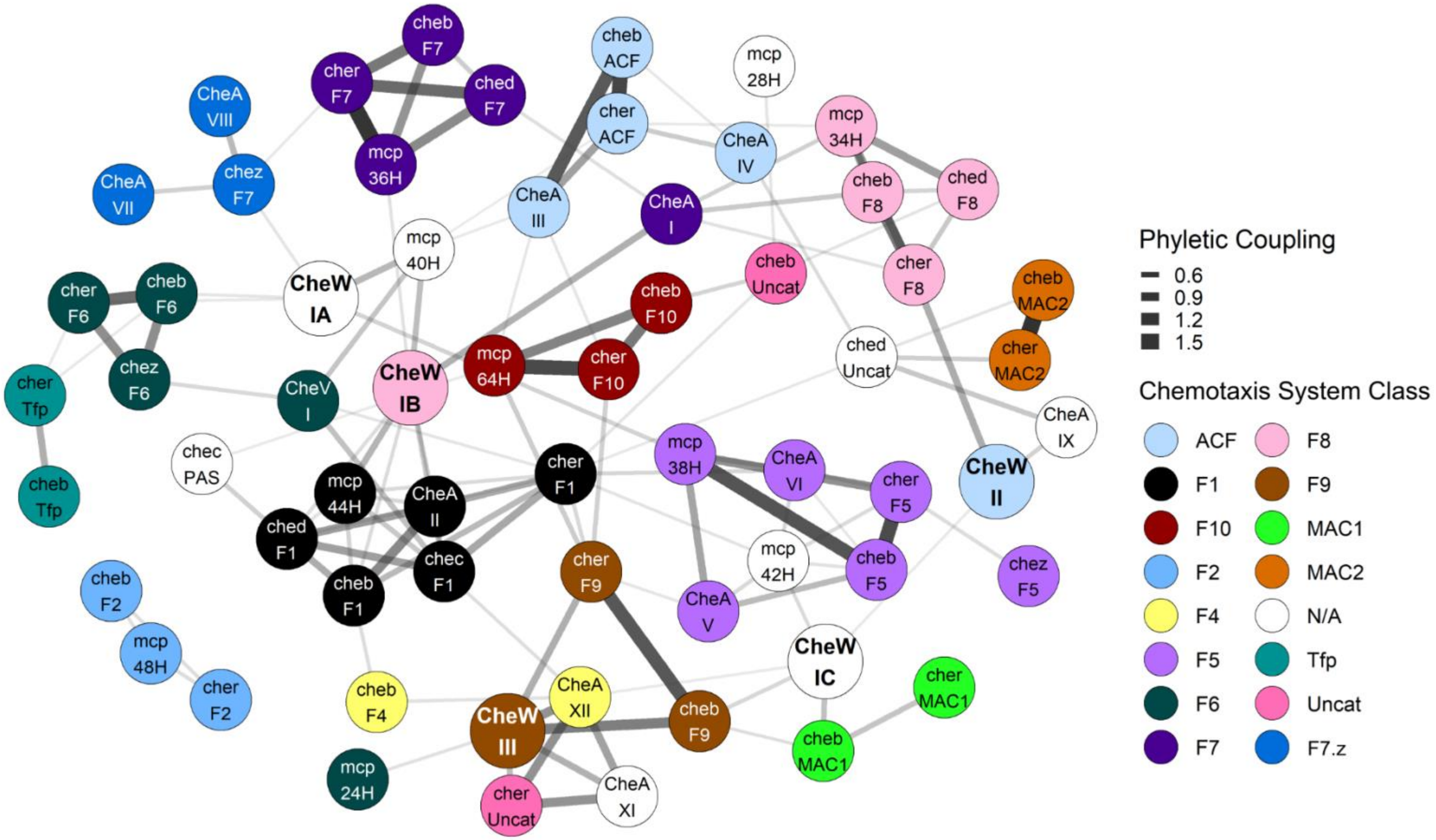


FIGURE 5. Network representation of inferred phyletic couplings between *Architectures* containing CheW-like domains and remaining chemotaxis system components. Co-occurrence data of chemotaxis components extracted from the RP35 representative proteome set were converted to a binary phylogenetic profile matrix. Phyletic Direct Coupling Analysis (PhyDCA) was used to quantify the favorability (correlation) of chemotaxis components co-occurring within the same organism. Strong favorability/high coupling typically corresponds to a cellular function (i.e., chemotaxis)

requiring both components, though not necessarily to a direct biophysical interaction. The top 125 positive co-evolutionary pairings were used to construct a graph based on phyletic coupling strength. Architectural assignments correspond to those included in Figure 1 (i.e., identical thresholds). Edge width was scaled with phyletic coupling strength. Notes: *Architecture* CheA.X did not appear in the top 125 strongest phyletic couplings and was excluded from the graph. cheb.F2, cher.F2, and mcp.48H formed a cluster disconnected from the rest of the network, but all three coupled to CheA.II at slightly lower strengths (top ~10%) (Dataset S4).