Revised for *Proteins: Structure, Function, and Bioinformatics*

5/31/22

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 Supporting Information 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 or 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 and form 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, revealing new insights. Co-occurrence analyses suggested that many chemotaxis systems occur in non-random combinations within species, implying synergy or antagonism. Furthermore, many *Architectures* of proteins containing CheW-like domains occurred predominantly with specific categories of chemotaxis systems, suggesting specialized functional interactions. We propose *Class* 1 (~80%) and *Class* 2 (~20%) CheW proteins exhibit preferences for distinct chemoreceptor structures. Furthermore, rare (~1%) Class 6 CheW proteins frequently co-occurred with methyl-accepting coiled coil (MAC) proteins, which contain both receptor and kinase functions and so do not require connection via a CheW scaffold but may benefit from arrays.

Lastly, rare multi-domain CheW proteins may interact with different receptors than single domain CheW proteins.

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,

78 including wide dynamic ranges, integration of mixed inputs, cooperativity, and rapid
79 signal amplification potential.[11] The sophisticated information processing capabilities of
80 the chemotaxis system are advantageous not only for general survival, but also for
81 invasion-, colonization-, and virulence-related processes in pathogenic
82 microorganisms.[12-16]

83 The chemotaxis pathway of *Escherichia coli* has been thoroughly characterized
84 and is an example of a two-component system that incorporates additional proteins to
85 achieve a more rapid and coordinated response.[17] The pathway begins at a
86 transmembrane array of chemoreceptors. Activation of the sensor array depends on
87 detection of an environmental stimulus and the methylation status of the receptors.
88 Following detection, a stimulus is converted into receptor conformational changes to
89 initiate processing and propagation. The signal is then passed to the histidine kinase,
90 CheA, which integrates information from multiple receptors through autophosphorylation
91 after summing positive and negative stimuli.[18,19] The signal path then splits into
92 “excitation” and “adaptation” branches. In the excitation path, phosphoryl groups are
93 passed to the response regulator CheY. Phosphorylation alters the equilibria between
94 active and inactive conformations in the CheY population, which ultimately modulates
95 flagellar motor behavior and motility. The adaptation path in *E. coli* features CheR and
96 CheB. CheR includes a methyltransferase domain that steadily adds methyl groups to
97 the chemoreceptors, independent of environmental stimuli. CheB is a response
98 regulator that includes a methylesterase domain whose activity is tightly regulated by
99 phosphorylation and removes methyl groups from the chemoreceptors in response to a
100 sufficient environmental change. The adaptation path forms a delayed negative

101 feedback loop, imparting a “memory” to the system and allowing the organism to both
102 follow a stimulus gradient and reset upon reaching a uniform environment. Some
103 chemotaxis systems also incorporate separate phosphatases, such as CheZ, to
104 catalyze the removal of phosphoryl groups and terminate the response at a specific
105 point in the pathway.[20] Many organisms encode multiple chemotaxis systems for
106 regulating multiple forms of propulsion and/or gene expression[5].

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

121 Most existing information on CheW-like domains describes the canonical,
122 standalone CheW protein (the most abundant form in nature).[25] The interactions

123 between the free species of CheW and its partners (MCPs and CheAs) have been well-
124 characterized for several microbial species, particularly *E. coli.*[26-30]
125 The most common occurrence of CheW-like domains other than CheA- or
126 CheW-lineage proteins (-lineage referring to proteins that can be characterized as
127 analogous to the canonical CheA and standalone CheW proteins in *E. coli*) is fused to a
128 receiver domain in CheV proteins (reviewed in [22,31]). Various commonly studied
129 organisms, including *Bacillus subtilis*, *Helicobacter pylori,* and *Vibrio cholerae,* encode
130 one or more CheV proteins. Although CheV proteins are present in approximately one-
131 third of all chemotaxis systems, their role(s) are still poorly understood.[32] CheV is
132 thought to be involved in both CheA modulation/MCP adaptation and array
133 formation/polar localization.[33,34] The receiver domain of CheV may also serve as a
134 general phosphate sink for the system.[31,35]
135 A landmark study by Wuichet and Zhulin established an evolutionary
136 classification of chemotaxis signaling systems in prokaryotes[32], summarized here. The
137 core components of essentially all chemotaxis systems, as outlined above, are the
138 MCP•CheW•CheA arrays, the CheB and CheR adaptation enzymes, and CheY
139 response regulators. Chemotaxis systems often have multiple MCPs and CheW
140 proteins, and it is technically difficult to distinguish CheY from other single domain
141 response regulators. Therefore, to provide a consistent foundation, the original
142 classification was based on phylogenetic trees of CheA/CheB/CheR sequences and
143 supported by certain other phylogenetic markers. There are 19 known categories of
144 standard chemotaxis systems, each containing characteristic arrangements of *che*
145 genes, distinct sets of auxiliary components (CheC phosphatase, CheD deamidase,

146 CheV scaffold protein, CheX phosphatase, and/or CheZ phosphatase), and often
147 unique architectures for certain core components. Seventeen categories of chemotaxis
148 systems (F1 through F17) control flagellar motility, one controls Type IV pili (Tfp), and
149 one controls alternative (non-motility) cellular functions (ACF). In addition, there are two
150 categories of related chemotaxis systems based on methyl-accepting coiled coil (MAC)
151 proteins, which contain both receptor and kinase functions, rather than separate MCP
152 and CheA proteins.

153 A recent companion paper describes our classification of CheW-like domains into
154 six *Classes*, likely related to specific functional specializations.[36] Nearly all (~94%)
155 CheW-like domains are encompassed by 16 distinct *Architectures* (Figure 1). CheW-like
156 domains in proteins with the CheW.I *Architecture* belong to three different *Classes*,
157 whereas CheW-like domains in other *Contexts* predominantly correspond to single
158 *Classes*. Most CheW-like domains in CheW- and CheV-lineage proteins belong to *Class*
159 1. Most CheW-like domains in CheA-lineage proteins belong to *Class* 3, except for the
160 CheA.VII and CheA.X *Architectures* (*Class* 4, which contain multiple Hpt domains) and
161 the *Class* 5 C-terminal CheW-like domains in CheA proteins with two such domains.
162 Rare (~1%) *Class* 2 CheW-like domains are found in CheW.I *Architectures* and exhibit
163 properties of both CheA-lineage and CheW-lineage proteins. About 20% of CheW-like
164 domains in CheW-lineage proteins (*Class* 6) appear subtly different from *Class* 1 and
165 may form yet another specialized *Class*. Strikingly, this study provides insights into both
166 Class 2 and Class 6 CheW-like domains.

167 In this work, we combined our classification scheme of CheA-, CheW-, and
168 CheV-lineage (CheW-like domain containing) proteins[36] with that of Wuichet and Zhulin

169 for other chemotaxis proteins[32] to gain additional insights into the occurrence and
170 organization of chemotaxis systems. We found that (i) many chemotaxis system
171 categories occur in non-random combinations within microbial species, and (ii) specific
172 *Architectures* of CheA/CheW proteins are preferentially associated with specific
173 chemotaxis system categories, suggesting functional interactions.

174 **2 | MATERIALS AND METHODS**

175 **2.1 | Protein sequence database and co-occurrence analysis**

176 Analysis of CheW-like domains (PF01584) was described in [36], using sequences
177 sourced from Representative Proteome 35 (RP35)[37] and the Pfam database (version
178 33, obtained May 2020).[38] We previously extracted all proteins that contained CheW-
179 like domains and belonged to the 16 *Architectures* (Figure 1) that account for ~94% of
180 CheW-like domains within the RP35 dataset.[36]

181 To analyze the co-occurrence (presence-absence) patterns of the various
182 chemotaxis components encoded by proteomes within the same dataset, proteins
183 containing one or more of the following domains were extracted: MCP (PF00015); CheR
184 (PF01739); CheB (PF01339); CheD (PF03975); CheZ (PF04344); CheCX (also simply
185 called CheC, PF04509/PF13690).[38] Receiver domain-containing proteins orthologous to
186 CheY were excluded for the sake of interpretability. Full protein sequences (excluding
187 the previously analyzed CheW-containing proteins) were scanned and classified with
188 HMMER3 (version 3.3) using the previously described chemotaxis system models
189 (utilized by the MiST database; version 3.0).[32,39-41] Thus, MCPs were classified by
190 number of heptad repeats,[42] whereas CheB, CheC (including closely related CheX[43]),
191 CheD, CheR, and CheZ proteins were assigned to the chemotaxis system categories of
192 Wuichet and Zhulin.[32] A combined frequency table with organisms in columns and
193 chemotaxis components in rows was generated by merging the previously classified
194 CheW-containing *Architectures* with the other chemotaxis components by organism.
195 Components with fewer than 20 positive occurrences were discarded from the analysis.
196 The excluded components belonged to eight categories (F3 or F11 through F17) of

197 chemotaxis systems, which were also previously observed to be infrequent.[32] The
198 resulting count matrix of distinct chemotaxis components encoded by 1887 distinct
199 proteomes (Dataset S1) was used to generate a heatmap (using the R package
200 ComplexHeatmap, version 2.7.11) of the co-occurrence patterns within the
201 representative proteomes.[44] The taxize package in R (version 0.9.99.947)[45] was used to
202 assign putative phyla and classes to the batch of relevant organisms (sourced from the
203 NCBI Taxonomy Browser[46]). Assignments were visualized as row/column annotations
204 with ComplexHeatmap. Rows (components) and columns (species) were grouped in an
205 unsupervised manner by hierarchical clustering, based on Spearman correlation, using
206 Ward's clustering method (option "ward.D2"). The resulting dendrograms (both row and
207 column) were split by height (using the cutree() function, as implemented by the
208 row_split/column_split options) into putative functional "blocs" of chemotaxis
209 components. The choice of height does not meaningfully affect the order of rows and
210 columns in the heatmap but does affect the number of blocs/splits displayed. We
211 selected a single height to cut the row dendogram such that components from the same
212 chemotaxis system categories largely ended up in the same row bloc (i.e. all F1
213 components together, all F6 components together, etc.), making strong biological
214 sense. The height to cut the column dendogram was chosen to optimize interpretability,
215 i.e., visual segregation of data into discrete column blocks
216 To analyze the co-occurrence patterns of the chemotaxis classes themselves,
217 rather than the individual components, the original matrix (Dataset S1) generated to
218 create Figure 2 was binarized. Rows featuring CheA, CheW, CheV and MCP paralogs
219 were first removed because they are used in multiple chemotaxis system categories. A

220 binary scheme was then applied for each individual chemotaxis system category (F1,
221 F2, F4, F5, F6, F7, F7.z, F8, F9, F10, Tfp, ACF, MAC1, MAC2 and Uncat). If a given
222 organism contained at least one of the corresponding components (CheB/C/D/R/Z) for a
223 chemotaxis category, then the category was considered present in the final table (=1),
224 irrespective of the presence/absence of other chemotaxis system categories (or
225 paralogous instances of the same category). Those categories lacking all components
226 were assigned absent (=0). A small percentage of organisms (<5%) lacked auxiliary
227 components entirely and were excluded. Dataset S2 provides a full breakdown with
228 corresponding proteome counts. A new heatmap (Figure 3) was generated in a similar
229 manner using ComplexHeatmap to re-cluster the proteomes (i.e., columns; 1797 distinct
230 organisms). Row order was maintained to match Figure 2. Dataset S3 provides the
231 relative abundances of the various chemotaxis system categories across all proteomes.

232 **2.2 | Phyletic direct coupling analysis (PhyDCA) and network generation**

233 The traditional approach to phylogenetic profiling utilizes some form of local
234 correlation metric, such as Hamming distance or Pearson correlation, to transform a
235 simple binary matrix representing presence (1) or absence (0) of specific
236 components/proteins/genes in various species into a corresponding interaction network.
237 However, classical profiling suffers from several disadvantages, such as the influence of
238 “intermediate” effects on apparent direct couplings (meaning that if A co-evolves with B,
239 and B co-evolves with C, A may also appear to co-evolve with C). A more recent
240 approach introduced the concept of direct coupling analysis, a statistical modeling
241 technique able to distinguish between direct and more indirect co-evolutionary signals,
242 to the profiling of presence-absence patterns.[47] This method, called Phyletic Direct

243 Coupling Analysis, or PhyDCA, has demonstrated substantially increased accuracy
244 compared to the more traditional correlation-based approaches and provided a
245 convenient means by which to quantify the relationships presented in Figure 2. The
246 frequencies used to create the Figure 2 co-occurrence heatmap (covering the full, non-
247 filtered complement of chemotaxis components) were converted into a binary
248 phylogenetic profile matrix to analyze pairwise evolutionary couplings. Data were
249 analyzed with PhyDCA (using the mfDCA implementation) to estimate relevant
250 quantitative phyletic pairings using a global statistical modelling approach.[47] The
251 phyletic coupling (J_{ij}) between two domains and/or components in our data was used to
252 estimate the favorability of finding multiple elements within the same species,
253 corresponding to the principle that a biological process (i.e., chemotaxis) would require
254 both components to function and produce a strong positive coupling. A negative
255 coupling could also be interpreted as alternative solutions for similar functionality in a
256 given system. We took the 125 (4%) strongest positive pairwise couplings (i.e., the
257 presence of one component favors the presence of the other) and created a non-
258 directed graph to visualize the web of co-evolutionary signals using the R packages
259 igraph and ggraph (using the Fruchterman-Reingold algorithm).[48-50] The phyletic
260 coupling scores of the pairs displayed in Figure 4 (≥ 0.44) were well within the range of
261 known significantly predictive scores for the PhyDCA method (as determined previously
262 by comparing with known pairwise relations as a reference set).[47] The complete list of
263 pairwise coupling strengths (>3,000 possible pairs) is in Dataset S4.

264 **3 | RESULTS**

265 **3.1 | Many combinations of chemotaxis system categories are non-randomly**
266 **distributed across prokaryotic species**

267 To better understand the organization of prokaryotic chemotaxis signal
268 transduction pathways, we first extracted all identifiable chemotaxis proteins except
269 CheYs from 1887 prokaryotic proteomes as described in Materials and Methods. Each
270 type of chemotaxis protein (CheA, CheB, etc.) was sorted as follows: MCPs were
271 classified by the number of heptad repeats,[42] proteins containing CheW-like domains
272 were classified by *Architecture* as in Figure 1, and all other chemotaxis proteins were
273 assigned to the chemotaxis system categories of Wuichet and Zhulin.[32] The end result
274 was 64 distinct chemotaxis components (Dataset S1). We then organized the
275 information in two dimensions, with species in columns and chemotaxis components in
276 rows. Hierarchical clustering was performed to optimally group organisms and
277 chemotaxis components with similar co-occurrence patterns. The results were
278 visualized in a composite heatmap, with color indicating the number of instances of
279 each component in a proteome (Figure 2).

280 Wuichet and Zhulin classified chemotaxis system categories primarily based on
281 phylogenetic trees of CheA, CheB, and CheR protein sequences.[32] The dominant
282 feature of Figure 2 is that when clustered in an unsupervised manner the occurrence
283 (not sequence) data primarily formed recognizable blocs as a function of putative
284 chemotaxis system category (row dendrogram shown in Figure S1) and proteome. Such
285 a phenomenon strongly suggested that the data were linked in both dimensions, across
286 chemotaxis system categories and across proteomes. If each species (proteome)

287 encoded a single category of chemotaxis system, then two-dimensional clustering
288 would be trivial, and proteomes would group perfectly into functional blocs by system
289 category. However, >50% of all prokaryotic genomes that encode chemotaxis systems
290 contain multiple systems (first determined in [32] and again corroborated by our work).
291 With many known categories of chemotaxis systems, if combinations featuring multiple
292 categories in a single species were random, then clustering by species would be
293 disrupted, precluding the previously described functional blocs. Therefore, we infer that
294 many of the naturally occurring combinations of chemotaxis system categories with
295 prokaryotic species are non-random. We explore in qualitative and quantitative detail
296 multiple aspects of the relationships between chemotaxis components, chemotaxis
297 system categories, and proteomes in the following sections.

298 **3.2 | Presence-absence analysis reveals preferred and disfavored category**
299 **combinations in organisms with multiple chemotaxis systems**

300 By converting the frequency table of components (Dataset S1) used to generate
301 Figure 2 into a simplified, binary presence/absence matrix featuring only the chemotaxis
302 system categories themselves, we next determined the most common naturally
303 occurring combinations of chemotaxis system categories (Dataset S2) and generated a
304 heatmap displaying the presence/absence of each chemotaxis system category across
305 species (Figure 3).

306 We first sought to compare the distribution of chemotaxis system categories
307 across species with the evolutionary relationships between the chemotaxis systems.
308 The phylogenetic tree of chemotaxis systems features three main branches (here
309 arbitrarily designated Branches 1, 2 and 3), with Branch 2 exhibiting three sub-

310 Branches.[32] The most common 10% of chemotaxis category combinations observed in
311 Dataset S2 accounted for two-thirds of the proteomes in our study and are displayed in
312 relation to the various Branches in Table 1, with cross-referencing to their locations in
313 Figure 3. Approximately a third of the proteomes encoded only a single category of
314 chemotaxis system (Table 1, top).

315 We next focused on the proteomes encoding multiple chemotaxis systems, first
316 examining pairwise combinations of categories found within the same Branch (Table 1,
317 middle). In general, flagellar chemotaxis system categories within the same Branch
318 (F1/F2, F4/F9/F10, and F7.z/F7/F8) co-occurred at substantially higher frequencies in
319 Dataset S2 than would be expected based on frequencies of the constituent categories
320 from Dataset S3 (see calculations in Table S1). Conversely, the combination of
321 categories F5/F6 occurred approximately five times less frequently than expected using
322 the same relationship, implying some form of negative selection. It seems plausible that
323 the components of the two categories (F5/F6) may interfere with one another (or share
324 overlapping functions to some degree). In contrast to the flagellar systems, categories
325 ACF/Tfp and MAC1/2 both co-occurred at frequencies consistent with a random
326 distribution.

327 Over 50% of all proteomes encoding multiple distinct chemotaxis systems in our
328 dataset included categories from disparate Branches of the classification tree (Table 1
329 bottom and Dataset S2), suggesting highly diverse origins for most systems. We found
330 that most outgroup combinations occurred at frequencies relatively consistent with a
331 random distribution (Table S2). A notable exception was class F7.z, which occurred with
332 categories in outgroup Branches much less frequently than expected (consistent with a

333 strong/semi-exclusive linkage between F7.z/F7 and F7.z/F8). Additionally, class F9 co-
334 occurred with both F5 and F8 systems more frequently than expected from a random
335 distribution.

336 Finally, we examined the co-occurrent relationships between the flagellar
337 chemotaxis system classes and the ACF/Tfp/MAC1/MAC2 systems. *A priori*, we
338 speculated that the non-flagellar systems would operate independently from the
339 flagellar-controlling classes, revealing no obvious selective pressure(s). While the
340 majority of pairwise combinations between flagellar and non-flagellar systems
341 (approximately 67%) supported our prediction, a full third deviated substantially (Table
342 S3). One-quarter of pairwise combinations were observed at lower-than-expected
343 frequencies, with nearly half of the cases of negative selection involving either F1 or F2
344 systems. Ten percent of pairwise combinations were observed at higher-than-expected
345 frequencies, with nearly half of the cases of positive selection involving F10 systems.

346 **3.3 | Matching *Architectures* of proteins containing CheW-like domains to**
347 **preferred chemotaxis system categories**

348 Examining the functional blocs revealed by clustering in Figure 2 suggested that
349 various *Architectures* of CheA and CheW proteins are differentially favored by divergent
350 chemotaxis systems. In particular, many *Architectures* clustered with components
351 belonging to one chemotaxis system category. Wuichet and Zhulin described eight
352 cases of distinct architectures for proteins containing CheW-like domains that were
353 characteristic of specific chemotaxis system categories.[32] The four *Architectures* noted
354 by Wuichet and Zhulin that were sufficiently abundant to be included in our study were
355 CheA.III/CheA.IV, CheA.VI, CheA.XII, and CheW.III, which were linked to categories

356 ACF/F3, F5, F4, and F9, respectively. Our data confirmed most previous assignments.
357 Our larger sample size enabled us to also propose multiple additional assignments.
358 Qualitative assignments of specific *Architectures* to specific chemotaxis systems
359 inferred from sorting patterns in Figure 2 are summarized in Table 2. Quantitative
360 analyses reported in Section 3.4 strengthen and extend these observations. The 17 new
361 relationships identified in our work link *Architectures* CheA.I, CheA.II, CheA.V,
362 CheA.VII, CheA.VIII, CheV.I, CheW.IA, CheW.IB, CheW.IC, and CheW.II with
363 chemotaxis system categories F7/F8, F1, F5, F7.z, F7.z, F1/F6, F5/F6/F7.z, F1/F7/F8,
364 MAC1, and F8/ACF, respectively.

365 The information presented by Figure 2 is challengingly dense, but salient
366 features can be identified and discussed most easily using a grid coordinate system in
367 which blocs are identified by assigned chemotaxis system category (rows, containing
368 individual protein components; labeled on left with silver boxes) and representative
369 proteome cluster (columns, containing distinct organisms; labeled on bottom with
370 numbers as proteome clusters). Several individual rows in Figure 2 exhibited
371 distributions from which we could glean additional insights. Prominent components
372 spanning numerous proteome clusters included mcp.44H, mcp.24H, mcp.40H,
373 mcp.34H, and mcp.36H (found in F1, F6, N/A, F8, and F7 blocs respectively); CheW.IB
374 (F8 blocs); CheW.IA, which includes most CheW proteins (N/A blocs above F8); and
375 CheA.I, the simplest and most common CheA *Architecture* (F7 blocs). The listed
376 components are known to constitute the core signaling pathway shared by all
377 chemotaxis systems. CheV.I (the sole version of CheV detected in any significant
378 abundance; see [36]), found in nearly a third of all chemotaxis systems in nature, also

379 spanned many proteome clusters (F6 blocs). Strikingly, the occurrence of CheV.I
380 correlated well with the mcp.40H type of chemoreceptor, strongly suggesting
381 preferential interaction(s) (N/A blocs above F8).

382 Figure 2 also revealed several key features shared by the MAC1/2 chemotaxis
383 categories. Methyl-accepting coiled-coil (MAC) proteins are closely related to
384 chemotaxis proteins but, to the best of our knowledge, have not been experimentally
385 characterized in any respect. MAC proteins include apparent chemoreceptor and kinase
386 domains, and either incorporate (MAC1) or are associated with (MAC2) CheB and
387 CheR related domains.[32] It is unclear whether MAC proteins are evolutionary precursors
388 of canonical chemotaxis systems or degenerate remnants. Proteome clusters 16, 17,
389 and 18 featured high concentrations of organisms encoding MAC1 and/or MAC2
390 components but seemingly lacked other types of chemotaxis systems (MAC1 and
391 MAC2 blocs). MAC systems were also scattered across numerous proteome clusters in
392 Figure 2, rather than remaining constrained to contiguous blocs. Such a distribution
393 suggested substantial phylogenetic prolificacy. Wuichet and Zhulin found that ~80% of
394 species with MAC systems encode additional chemotaxis systems.[32] The distribution of
395 MAC systems seen in Figure 3, based on our much larger sample size, supported and
396 strengthened the original observation.

397 **3.4 | Co-evolutionary analysis reveals functional communities of chemotaxis**
398 **components**

399 Due to the complex nature of the information represented in Figure 2 (and the
400 relative inability of the human eye to untangle multivariate correlations), we sought a
401 way to simplify and quantify the co-occurrence probabilities of individual components

402 observed in the various chemotaxis systems in nature. We used Phyletic Direct
403 Coupling Analysis (PhyDCA) as described in Materials and Methods to quantify pairwise
404 phyletic couplings between chemotaxis components and displayed the 4% strongest
405 positive couplings (for which the presence of one component accurately predicts the
406 presence of the other) in Figure 4.

407 The relationships revealed by Figure 4 largely corroborated the clustering
408 patterns of Figure 2 and the assignments made in Table 2. Individual chemotaxis
409 components typically associated closely with others of the same system category
410 (represented by shared colors in Figure 4), but less strongly to nodes outside the same
411 group. The network representation was particularly useful for visualizing relationships
412 between the disparate categories/chemoreceptors and the CheW-/CheV-lineage
413 *Architectures*. Most of the chemotaxis categories could be traced to a least one CheW-
414 lineage and CheA-lineage component in relatively short order. Category F2 components
415 and mcp.48H appeared as a cluster unconnected to other components in Figure 4, but
416 all exhibited couplings in the top ~10% to CheA.II (Dataset S4) and hence linked to the
417 F1 chemotaxis system.

418 Figure 2 shows four groups of components (labeled N/A or uncategorized) that
419 sorted into isolated blocs rather than associating with the standard chemotaxis system
420 categories. Figure 4 suggests that the components were not associated with one
421 another through co-evolutionary processes, but rather were dispersed and associated
422 with a diverse range of other proteins. One explanation for the differences between the
423 results in Figures 2 and 4 is correlation of individual components with multiple

424 chemotaxis system categories. Such a phenomenon would likely facilitate linkage in
425 Figure 4 but confound the clustering procedure used for Figure 2.
426 The strong connections between components observed in Figure 4 allowed us to
427 confirm many of the chemotaxis system class assignments proposed in Table 2 (based
428 on observations from Figure 2), as well as to infer several additional novel assignments.
429 In particular, inspection of Figure 4 provides insights about how various *Architectures*
430 and *Classes* of CheW and CheV scaffold proteins assort across the various chemotaxis
431 system categories. Insights into the minor CheW.IA and CheW.IC *Architectures*, as well
432 as CheV.I, can be found in the Discussion. Additional comments on CheA.IB, CheW.II,
433 and CheW.III *Architectures* are included in Text S1.

434 **3.5 | Negative phyletic couplings reveal putative overlapping functionality among**
435 **specific chemotaxis components**

436 The PhyDCA model can also be used to predict negative phyletic couplings, i.e.,
437 the presence of one component in a proteome disfavors the presence of another.[47]
438 Logic suggests that components in such a scenario likely share overlapping (or at least
439 closely related) functionalities (sometimes referred to as “alternative” solutions). For
440 example, the top negative coupling within Dataset S4 involved the components cher.F8
441 and cher.Uncat, presumably because both methyltransferases serve highly similar
442 functions. The third strongest negative coupling involved CheW.II and CheW.III.
443 Curiously, few other instances of anticorrelated components with presumably similar
444 functions are present in the list of top negative phyletic pairs, with most entries involving
445 disparate component types (and are therefore not likely to be consequences of
446 convergent evolution). The only exceptions (from the top 4% strongest anticorrelated

447 phyletic pairs) were CheW.IA with CheW.IC (both CheW-lineage scaffolds), chec.F1
448 with chez.F7 (both phosphatases), CheA.I with CheA.II (the two most abundant CheA-
449 lineage kinases), cher.F10 with cher.F5 (both methyltransferases), cheb.F1 with
450 cheb.F10 (both methylesterases), cher.F1 with cher.F8 (both methyltransferases),
451 cheb.F7 with cheb.F8 (both methylesterases), and finally cheb.F10 with cheb.F5 (both
452 methylesterases).
453

454 **4 | DISCUSSION**

455 **4.1 | Many combinations of chemotaxis system categories within a species are**
456 **non-random**

457 The observations described in this report, and the conclusions made or inferred,
458 were only possible because we sampled proteomes from a large number of distinct
459 organisms. The results of unsupervised two-dimensional sorting of species according to
460 their content of chemotaxis system categories (Figures 2 and 3) strongly implies that the
461 distribution of systems is substantially non-random. It is conceptually simple to sort
462 species encoding only one category into a bloc (e.g., F1 bloc in Column 1 of Figure 2 or
463 Column 6 of Figure 3). However, most species in our dataset encoded more than one
464 chemotaxis system. Non-random combinations of systems would preclude two-
465 dimensional sorting into blocs. The bloc data structure observed in Figures 2 and 3
466 suggested that a restricted subset of combinations of chemotaxis systems is
467 evolutionarily favored. Preferred combinations must be either ancient (passed on to
468 descendants of common ancestors) and/or are synergistically beneficial (arose
469 independently multiple times). In either case, horizontal gene transfer of chemotaxis
470 systems between species has not erased the pattern of combinatorial preferences in
471 nature. Though members of some prokaryotic phyla tend to encode particular
472 categories of chemotaxis systems, the topologies of the chemotaxis and species
473 classification trees do not match[32,51], implying different evolutionary paths. The primary
474 combinations of chemotaxis system categories (Table 1) and their non-random nature
475 (Figures 2 and 3) may provide clues into the rise of various chemotaxis systems.
476

477 Figures 2 and 3 revealed large scale qualitative features of the data. Closer
478 quantitative evaluation (Tables S1-S3) revealed more complex features. Many
479 combinations of systems co-occurred more or less frequently than expected (i.e., a non-
480 random pattern). Overall, our findings imply that most flagellar chemotaxis systems are
481 not independent from one another. This would be expected if multiple systems control
482 the same flagellar motors (e.g., having closely related CheY proteins to control the
483 same motor could be advantageous). In contrast, ACF and Tfp systems appeared to act
484 independently from each other, as did MAC1 and MAC2 systems.

485 **4.2 | Apparent functional specialization of proteins containing CheW-like domains**

486 CheW-like domains primarily occur in CheA, CheV, and CheW proteins, but the
487 CheA- and CheW-lineages exhibit substantial diversity (Figure 1).[36] Figure 2 shows that
488 CheA.I (the most abundant CheA *Architecture*), CheW.IA and CheW.IB (which together
489 comprise almost all CheW proteins), CheV.I (the sole abundant CheV *Architecture*), and
490 many types of MCPs were widely distributed across proteomes, implying an ability to
491 function in multiple chemotaxis systems. However, our analyses (summarized in Table
492 2) strongly suggest that many *Architectures* of proteins containing CheW-like domains
493 are associated with particular chemotaxis system categories, implying functional
494 specialization. We confirmed four assignments made by Wuichet and Zhulin[32] and
495 made 17 new assignments. The additional assignments provide a rich foundation for
496 future mechanistic investigations.

497 **4.3 | Insights into functions of uncommon Class 2 and Class 6 CheW-like domains**

498 Most CheW-like domains in CheW proteins belong to *Class* 1 (Figure 1, Ref. [36]).
499 Co-occurrence data provide potential insight into proteins with CheW.I Architecture but
500 less common *Class* 2 and *Class* 6 domains.

501 CheW.IA comprises *Class* 6 of CheW-like domains and makes up ~20% of
502 CheW.I proteins.[36] CheW.IA is subtly distinguishable from CheW.IB and CheV.I CheW-
503 like domains (*Class* 1) by some (but not all) methods of sequence analysis.[36] In Figure
504 4, CheW.IA made direct connections to various chemoreceptors, but not to any CheA
505 proteins (i.e., phyletic coupling of CheW.IA was stronger to MCPs than to CheA-lineage
506 proteins). We speculate that the distinction between Class 1 and Class 6 CheW-like
507 domains is that the latter exhibit greater specificity or preference for interactions with
508 certain classes of MCPs (e.g., mcp.40H, mcp.38H). Note that CheW.IA sorted with
509 mcp.40H in Figure 2 (N/A bloc above F8), but not with a specific chemotaxis system
510 class. In a related observation, CheW.IA and CheW.IB account for nearly all single
511 domain CheW proteins in nature. Both made strong direct connections to mcp.40H in
512 Figure 4. CheW.IB also made direct connections with mcp.44H (strong) and mcp.36H
513 (weaker), and CheW.IA made a direct connection (strong) with mcp.38H. Collectively,
514 MCPs from 36H, 38H, 40H and 44H account for nearly 90% of all MCPs in nature (at
515 least as encompassed by the RP35 dataset). When viewed in this way, the phyletic
516 couplings between such prolific components makes sense. However, we again
517 speculate that the “unique” interactions noted for CheW.IA and CheW.IB are rooted in
518 preferences for distinct chemoreceptors. Although both likely share a robust ability to
519 interact with mcp.40H, CheW.IA may also be able to interact with mcp.38H, whereas
520 CheW.IB may be able to interact with both mcp.44H and mcp.36H.

521 CheW.IA was directly linked to F5, F6, and F7.z chemotaxis system categories in
522 Figure 4. CheW.IB was directly linked to F1 and F7 components in Figure 4 and sorted
523 with F8 in Figure 2. As the F7.z category diverged from F7 and became associated with
524 CheA.VII and CheA.VIII rather than CheA.I *Architectures* (Table 2), CheW may have
525 diverged in parallel from *Class* 1 CheW.IB (F7) to *Class* 6 CheW.1A (F7.z).
526 CheW.IC (*Class* 2) makes up ~1% of CheW-like domains and shares
527 characteristics of CheW-like domains found in both CheA and CheW proteins.[36]
528 CheW.IC did not sort with any known chemotaxis system category in Figure 2 (N/A bloc
529 below F4), perhaps due to its low abundance in nature. The same rarity makes
530 interpretation of the distribution of CheW.IC in Figure 2 challenging. However, CheW.IC
531 was strongly linked with the MAC1 category in Figure 4 (corroborated upon close
532 inspection of Figure 2). MAC systems incorporate both receptor and kinase functions
533 into a single protein species, implying that they do not need CheW scaffold proteins to
534 bridge the two elements. It is not known if any MAC proteins form arrays in conjunction
535 with CheW proteins. In principle, arrays of MAC proteins could provide previously
536 described advantages (e.g., sensitive signal detection, amplification, integration) of a
537 canonical chemoreceptor array, but array properties might be constrained by the one-to-
538 one relationship between intramolecular chemoreceptor and kinase functions in MAC
539 proteins. Arrays could also facilitate adaptation, for example by allowing the CheB
540 and/or CheR domains of MAC1 proteins to modify adjacent receptors, or the separate
541 CheR proteins of MAC2 systems to localize to the array by molecular brachiation.[52]

542 **4.4 | CheW.II and CheW.III *Architectures* may interact with unusual MCPs**

543 CheW.II and CheW.III *Architectures* are ~1 to 3% as abundant as CheW.I
544 *Architectures*.[36] We sought to use negative PhyDCA pairings to infer the extent of
545 overlapping roles between CheW proteins with various *Architectures*. Several results
546 are consistent with the hypothesis that CheW.II and CheW.III are specialized to support
547 interactions with particular MCPs. First, the third strongest negative PhyDCA coupling
548 was between CheW.II and CheW.III (Dataset S4). A negative PhyDCA coupling
549 suggests that CheW.II and CheW.III perform similar functions, so only one is needed.
550 (An alternative possibility is that because CheW.II and CheW.III proteins are both rare,
551 our sample size is too small to include species encoding both). Second, we are aware
552 of only one experimental investigation of CheW.II or CheW.III function, which showed
553 that CheW.III in *Vibrio cholerae* linked an unusual type of MCP to CheA in cytoplasmic
554 double-layered arrays.[53] Finally, none of the top negative PhyDCA pairings involving
555 CheW.II or CheW.III featured any other CheW-lineage *Architecture*, implying, along with
556 frequent co-occurrences with other CheW proteins in Figure 2, that the CheW.II and
557 CheW.III *Architectures* are not "alternative" solutions for the more standard CheW-
558 lineage components. In other words, CheW.II probably does not replace two distinct
559 single-domain CheW proteins, but likely serves novel functionalities. However, an
560 apparently distinct role is constrained by the fact that, like the abundant CheW.IB,
561 CheW.II and CheW.III contain Class 1 CheW-like domains (Figure 1, Ref. [36]).

562 **4.5 | A potential link between CheV and CheC/CheZ phosphatases**

563 CheV.I clustered with category F6 chemotaxis systems in Figure 2 but also
564 appeared coincident with the F1 category. Figure 4 confirmed connections between
565 CheV.I and the F1 and F6 categories. Figure 2 also showed a strong correlation

566 between CheV.I and mcp.40H, which was again confirmed by Figure 4. However,
567 CheV.I did not show a direct connection to any specific CheA *Architecture*. Curiously,
568 besides a link with mcp.40H, the only other strong direct correlations formed by CheV.I
569 involved the phosphatases chez.F6/chec.F1 and the methyltransferase cher.F1. The
570 role(s) of CheV-lineage proteins and their attached receiver domains are poorly
571 understood. Some evidence suggests that CheV is involved in the chemotaxis
572 adaptation process[31], making the correlation between CheV.I and the CheC/CheR
573 components of the F1 class[31] understandable. However, the nature of the connection
574 between CheV.I and chez.F6 is less clear and raises the concept of CheZ (and possibly
575 the CheC of category F1) acting upon the attached phosphorylatable receiver domain of
576 CheV.I in the capacity of a phosphatase. In fact, CheZ has phosphatase activity toward
577 one of the three CheV proteins in *H. pylori*.[54] It is not known whether CheZ distinguishes
578 between different CheV proteins based on their CheW-like and/or receiver domains.

579 **4.6 | Limitations of PhyDCA**

580 The PhyDCA approach used to generate Figure 4 has several disadvantages
581 that must be considered. One is that the observed phyletic couplings do not necessarily
582 correspond to direct biophysical interactions. Strong coupling may also represent events
583 such as genomic co-localization (a limitation the original authors circumvented by
584 including an additional residue-level covariance analysis to predict likely direct
585 interaction partners).[47] Because of the narrow perspective of our study (i.e., focusing on
586 chemotaxis systems, rather than entire proteomes), we believe that this disadvantage
587 was minimized. However, as can be seen in Figure 2, paralogous chemotaxis
588 components are common in bacteria, particularly for chemoreceptors. Introducing a

589 residue-level analysis step may facilitate the untangling of specific paralog
590 interactions,[47,55] especially matching the various CheW-like domains in CheA-, CheW-,
591 and CheV-lineage proteins to their partner MCP components. Such an analysis would
592 provide additional insight into the diverse chemotaxis systems found in nature.
593 Additionally, PhyDCA (and many other phylogenetic profiling methods) relies on a
594 binary presence/absence data structure to simplify data processing and interpretation.
595 Excluding the substantial amount of paralogous protein data found in our microbial
596 dataset very likely ignores valuable co-occurrence information. However, our two-
597 pronged approach to the problem (using the full co-occurrence matrix to identify
598 functional blocs/clusters in Figure 2 and using the transformed binary profile to create
599 the simplified network representation in Figure 4) likely mitigates the issue.

600 **4.7 | General insights into occurrence and organization of prokaryotic chemotaxis**
601 **systems**

602 CheW-like domains play a central role in the signal transduction systems that
603 regulate prokaryotic chemotaxis by linking receptors and kinases into large arrays.
604 Almost all CheW-like domains occur in a limited number of *Architectures* of CheA-,
605 CheW-, and CheV-lineage proteins (Figure 1).[36] Furthermore, CheW-like domains have
606 evolved into distinct functional *Classes* (Figure 1).[36] We surveyed chemotaxis proteins
607 encoded by nearly 1900 species (Dataset S1) and examined their distribution by both
608 chemotaxis system category[32] and by species/proteome. Successful unsupervised
609 clustering of components into blocs (Figure 2) strongly suggested that the components
610 were linked in both dimensions, leading to two central conclusions. First, many
611 combinations of chemotaxis systems encoded by individual species (Table 1, Dataset

612 S2) were non-random (Figure 3, Tables S1-S3). Specific co-occurrence patterns and
613 frequencies (Dataset S3) should provide insights into evolution of chemotaxis systems.
614 Second, we inferred probable functional associations between each *Architecture* of
615 CheA-, CheW-, and CheV-lineage proteins and specific categories of chemotaxis
616 systems (Figure 4, Table 2, Dataset S4). These assignments lay a foundation for future
617 investigations into the mechanisms that underly apparent functional specialization of
618 different chemotaxis protein *Architectures*.
619

620 **ACKNOWLEDGEMENTS**

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

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

627 **REFERENCES**

628 1. Alvarez AF, Barba-Ostria C, Silva-Jiménez H, Georgellis D. Organization and mode
629 of action of two component system signaling circuits from the various kingdoms of
630 life. *Environ Microbiol.* 2016;18(10):3210-3226.
631 2. Zschiedrich CP, Keidel V, Szurmant H. Molecular mechanisms of two-component
632 signal transduction. *J Mol Biol.* 2016;428(19):3752-3775.
633 3. Gao R, Stock AM. Biological insights from structures of two-component proteins.
634 *Annu Rev Microbiol.* 2009;63(1):133-154.
635 4. Bourret RB, Kennedy EN, Foster CA, Sepúlveda VE, Goldman WE. A radical
636 reimagining of fungal two-component regulatory systems. *Trends Microbiol.*
637 2021;29(10):883-893.
638 5. Kirby JR. Chemotaxis-like regulatory systems: Unique roles in diverse bacteria.
639 *Annu Rev Microbiol.* 2009;63(1):45-59.
640 6. Popp F, Armitage JP, Schüler D. Polarity of bacterial magnetotaxis is controlled by
641 aerotaxis through a common sensory pathway. *Nat Commun.* 2014;5(1):5398.
642 7. Tohidifar P, Plutz MJ, Ordal GW, Rao CV. The mechanism of bidirectional pH taxis
643 in *Bacillus subtilis*. *J Bacteriol.* 2020;202(4).
644 8. Vaknin A, Berg HC. Osmotic stress mechanically perturbs chemoreceptors in
645 *Escherichia coli*. *Proc Natl Acad Sci U S A.* 2006;103(3):592-596.
646 9. Yang Y, Sourjik V. Opposite responses by different chemoreceptors set a tunable
647 preference point in *Escherichia coli* pH taxis. *Mol Microbiol.* 2012;86(6):1482-1489.
648 10. Yoney A, Salman H. Precision and variability in bacterial temperature sensing.
649 *Biophy J.* 2015;108(10):2427-2436.

650 11. Parkinson JS, Hazelbauer GL, Falke JJ. Signaling and sensory adaptation in
651 *Escherichia coli* chemoreceptors: 2015 update. *Trends Microbiol.* 2015;23(5):257-
652 266.
653 12. Yao J, Allen C. Chemotaxis is required for virulence and competitive fitness of the
654 bacterial wilt pathogen *Ralstonia solanacearum*. *J Bacteriol.* 2006;188(10):3697-
655 3708.
656 13. Butler SM, Camilli A. Going against the grain: Chemotaxis and infection in *Vibrio*
657 *cholerae*. *Nat Rev Microbiol.* 2005;3(8):611-620.
658 14. Chandrashekhar K, Kassem II, Rajashekara G. *Campylobacter jejuni* transducer like
659 proteins: Chemotaxis and beyond. *Gut Microbes.* 2017;8(4):323-334.
660 15. Josenhans C, Suerbaum S. The role of motility as a virulence factor in bacteria. *Int J*
661 *Med Microbiol.* 2002;291(8):605-614.
662 16. Kreling V, Falcone FH, Kehrenberg C, Hensel A. *Campylobacter sp.*: Pathogenicity
663 factors and prevention methods—new molecular targets for innovative antivirulence
664 drugs? *Appl Microbiol Biotechnol.* 2020;104(24):10409-10436.
665 17. Sourjik V, Wingreen NS. Responding to chemical gradients: Bacterial chemotaxis.
666 *Curr Opin Cell Biol.* 2012;24(2):262-268.
667 18. Khan S, Spudich JL, McCray JA, Trentham DR. Chemotactic signal integration in
668 bacteria. *Proc Natl Acad Sci U S A.* 1995;92(21):9757-9761.
669 19. Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA. The two-component
670 signaling pathway of bacterial chemotaxis: a molecular view of signal transduction
671 by receptors, kinases, and adaptation enzymes. *Annu Rev Cell Dev Biol.*
672 1997;13:457-512.

673 20. Zhao R, Collins EJ, Bourret RB, Silversmith RE. Structure and catalytic mechanism
674 of the *E. coli* chemotaxis phosphatase CheZ. *Nat Struct Biol.* 2002;9(8):570-575.
675 21. Bi S, Sourjik V. Stimulus sensing and signal processing in bacterial chemotaxis. *Curr*
676 *Opin Microbiol.* 2018;45:22-29.
677 22. Huang Z, Pan X, Xu N, Guo M. Bacterial chemotaxis coupling protein: Structure,
678 function and diversity. *Microbiol Res.* 2019;219:40-48.
679 23. Pinas GE, DeSantis MD, Cassidy CK, Parkinson JS. Hexameric rings of the
680 scaffolding protein CheW enhance response sensitivity and cooperativity in
681 *Escherichia coli* chemoreceptor arrays. *Sci Signal.* 2022;15(718):eabj1737.
682 24. Bilwes AM, Alex LA, Crane BR, Simon MI. Structure of CheA, a signal-transducing
683 histidine kinase. *Cell.* 1999;96(1):131-141.
684 25. Park S-Y, Borbat PP, Gonzalez-Bonet G, et al. Reconstruction of the chemotaxis
685 receptor–kinase assembly. *Nat Struct Mol Biol.* 2006;13(5):400-407.
686 26. Ortega DR, Mo G, Lee K, et al. Conformational coupling between receptor and
687 kinase binding sites through a conserved salt bridge in a signaling complex scaffold
688 protein. *PLOS Comp Biol.* 2013;9(11):e1003337.
689 27. Wang X, Vu A, Lee K, Dahlquist FW. CheA-receptor interaction sites in bacterial
690 chemotaxis. *J Mol Biol.* 2012;422(2):282-290.
691 28. Boukhvalova MS, Dahlquist FW, Stewart RC. CheW binding interactions with CheA
692 and Tar: Importance for chemotaxis signaling in *Escherichia coli*. *J Biol Chem.*
693 2002;277(25):22251-22259.

694 29. Bellenger K, Ma X, Shi W, Yang Z. A CheW homologue is required for *Myxococcus*
695 *xanthus* fruiting body development, social gliding motility, and fibril biogenesis. *J*
696 *Bacteriol.* 2002;184(20):5654-5660.
697 30. Martin AC, Wadhams GH, Armitage JP. The roles of the multiple CheW and CheA
698 homologues in chemotaxis and in chemoreceptor localization in *Rhodobacter*
699 *sphaeroides*. *Mol Microbiol.* 2001;40(6):1261-1272.
700 31. Alexander RP, Lowenthal AC, Harshey RM, Ottemann KM. CheV: CheW-like
701 coupling proteins at the core of the chemotaxis signaling network. *Trends Microbiol.*
702 2010;18(11):494-503.
703 32. Wuichet K, Zhulin IB. Origins and diversification of a complex signal transduction
704 system in prokaryotes. *Sci Signal.* 2010;3(128):ra50-ra50.
705 33. Lowenthal AC, Simon C, Fair AS, et al. A fixed-time diffusion analysis method
706 determines that the three *cheV* genes of *Helicobacter pylori* differentially affect
707 motility. *Microbiology.* 2009;155(Pt 4):1181-1191.
708 34. Yang W, Alvarado A, Glatter T, Ringgaard S, Briegel A. Baseplate variability of
709 *Vibrio cholerae* chemoreceptor arrays. *Proc Natl Acad Sci U S A.*
710 2018;115(52):13365-13370.
711 35. Ortega DR, Zhulin IB. Evolutionary genomics suggests that CheV Is an additional
712 adaptor for accommodating specific chemoreceptors within the chemotaxis signaling
713 complex. *PLOS Comp Biol.* 2016;12(2):e1004723.
714 36. Vass LR, Bascum KM, Bourret RB, Foster CA. Generalized strategy to analyze
715 domains in the context of parent protein architecture: Case study of CheW. *Proteins.*
716 2022;Submitted.

717 37. Chen C, Natale DA, Finn RD, et al. Representative proteomes: A stable, scalable
718 and unbiased proteome set for sequence analysis and functional annotation. *PLoS*
719 *One.* 2011;6(4):e18910-e18910.
720 38. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in
721 2021. *Nucleic Acids Res.* 2020;49(D1):D412-D419.
722 39. Gumerov VM, Ortega DR, Adebali O, Ulrich LE, Zhulin IB. MiST 3.0: An updated
723 microbial signal transduction database with an emphasis on chemosensory systems.
724 *Nucleic Acids Res.* 2019;48(D1):D459-D464.
725 40. Wuichet K, Alexander RP, Zhulin IB. Comparative genomic and protein sequence
726 analyses of a complex system controlling bacterial chemotaxis. *Methods Enzymol.*
727 2007;422:1-31.
728 41. Eddy SR. Accelerated Profile HMM Searches. *PLOS Comp Biol.*
729 2011;7(10):e1002195.
730 42. Alexander RP, Zhulin IB. Evolutionary genomics reveals conserved structural
731 determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl*
732 *Acad Sci U S A.* 2007;104(8):2885-2890.
733 43. Silversmith RE. Auxiliary phosphatases in two-component signal transduction. *Curr*
734 *Opin Microbiol.* 2010;13(2):177-183.
735 44. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in
736 multidimensional genomic data. *Bioinformatics.* 2016;32(18):2847-2849.
737 45. Chamberlain SA, Szocs E. taxize: taxonomic search and retrieval in R. *F1000Res.*
738 2013;2:191.

739 46. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update
740 on curation, resources and tools. *Database (Oxford).* 2020;2020.
741 47. Croce G, Gueudré T, Ruiz Cuevas MV, et al. A multi-scale coevolutionary approach
742 to predict interactions between protein domains. *PLOS Comp Biol.*
743 2019;15(10):e1006891.
744 48. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement.
745 *Software Pract Exper.* 1991;21(11):1129-1164.
746 49. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph:
747 Network visualizations of relationships in psychometric data. *J Stat Softw.*
748 2012;48(4):1 - 18.
749 50. *ggraph: An implementation of grammar of graphics for graphs and networks*
750 [computer program]. Version 2.0.52021.
751 51. Gumerov VM, Andrianova EP, Zhulin IB. Diversity of bacterial chemosensory
752 systems. *Curr Opin Microbiol.* 2021;61:42-50.
753 52. Levin MD, Shimizu TS, Bray D. Binding and diffusion of CheR molecules within a
754 cluster of membrane receptors. *Biophys J.* 2002;82(4):1809-1817.
755 53. Briegel A, Ortega DR, Mann P, Kjaer A, Ringgaard S, Jensen GJ. Chemotaxis
756 cluster 1 proteins form cytoplasmic arrays in *Vibrio cholerae* and are stabilized by a
757 double signaling domain receptor DosM. *Proc Natl Acad Sci U S A.*
758 2016;113(37):10412-10417.
759 54. Lertsethtakarn P, Ottemann KM. A remote CheZ orthologue retains phosphatase
760 function. *Mol Microbiol.* 2010;77(1):225-235.

761 55. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous
762 identification of specifically interacting paralogs and interprotein contacts by direct
763 coupling analysis. *Proc Natl Acad Sci U S A.* 2016;113(43):12186-12191.
764

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 3
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[32] 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 2.

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

Protein *Architecture*[a]	CheW-like Domain *Class*[b]	Chemotaxis System Category[c]	Evidence for Chemotaxis System Assignment		
			Figure 6 of [32]	Figure 2	Figure 4
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 1. The CheW.I *Architecture* splits into three *Classes*.[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.

843 [d]Wuichet & Zhulin[32] noted CheA proteins modified only by C-terminal receiver domains (CheA.III or CheA.IV) were
844 consistently observed in F3 chemotaxis systems. However, our PhyDCA scores in Dataset S4 do not support linkage of
845 either the CheA.III or CheA.IV *Architectures* to either F3 or F4 chemotaxis systems.
846 [e]Although the CheA.V, CheA.VII, CheA.VIII, CheA.X, and CheW.II *Architectures* contain additional domains with respect
847 to canonical CheA or CheW *Architectures* and are sufficiently abundant to be included in [36], these *Architectures* were
848 apparently not observed sufficiently consistently in specific chemotaxis system categories to be noted by Wuichet &
849 Zhulin[32], who analyzed a much smaller sample size.
850 [f]The subset of F7 chemotaxis systems that contain CheZ. See Figure 2.
851 [g]Did not sort with a specific chemotaxis system category in Figure 2 and did not make sufficiently strong connections to be
852 included in Figure 4.
853

854 **FIGURE LEGENDS**

855 **FIGURE 1. Major *Architectures* of proteins that contain CheW-like domains** (from
856 [36]). CheA-lineage *Architectures* are designated by a Roman numeral suffix in order of
857 decreasing abundance. CheW-lineage *Architectures* are designated by a Roman
858 numeral suffix indicating the number of CheW-like domains. The CheW.I *Architecture*
859 includes sequences that belong to three distinct *Classes*, designated CheW.IA,
860 CheW.IB, and CheW.IC. *Classes* of CheW-like domains are indicated by white
861 numerals.

862 **FIGURE 2. Co-occurrences of individual chemotaxis components in RP35**
863 **representative proteome set.** Total occurrence counts were used. Components with <
864 20 occurrences were excluded. Column annotations were shaded by Phyla (groups with
865 < 10 occurrences are unlabeled) and Class (groups with < 10 occurrences are
866 unlabeled). Notable organisms were tagged. Results were split by dendrogram height
867 into functional "blocs" by clustering both proteomes (columns) and chemotaxis
868 components (rows). Representative Proteome clusters were labeled as 1-18, whereas
869 component blocs were labeled with most likely chemotaxis system category. Row
870 dendogram shown in Figure S1.

871 **FIGURE 3. Simplified co-occurrence schematic of chemotaxis system categories**
872 **in RP35 representative proteome set.** A binary presence/absence scheme was used
873 for visualization. A chemotaxis system category was determined to be present in a
874 given proteome if at least one of the following components was detected of the
875 appropriate category: CheB/C/D/R/Z. CheA/V/W and MCP components were excluded

876 from the analysis, because some of these components function with more than one
877 chemotaxis system category. Notable organisms were labelled. Row order was
878 maintained for consistency with Figure 2, meaning categories were not clustered.
879 Results were split by dendrogram height into functional “blocs” by clustering proteomes
880 (columns). However, because the datasets upon which Figures 2 and 3 are based are
881 different, the resulting proteome clusters and cluster numbers are different than in
882 Figure 2.

883 **FIGURE 4. Network representation of inferred phyletic couplings between**
884 ***Architectures* containing CheW-like domains and remaining chemotaxis system**
885 **components.** Co-occurrence data of chemotaxis components extracted from the RP35
886 representative proteome set were converted to a binary phylogenetic profile matrix.
887 Phyletic Direct Coupling Analysis (PhyDCA) was used to quantify the favorability
888 (correlation) of chemotaxis components co-occurring within the same organism. Strong
889 favorability/high coupling typically corresponds to a cellular function (i.e., chemotaxis)
890 requiring both components, though not necessarily to a direct biophysical interaction.
891 The top 125 (~4%) positive co-evolutionary pairings were used to construct a graph
892 based on phyletic coupling strength. Architectural assignments correspond to those
893 included in Figure 1 (i.e., identical thresholds). Edge width was scaled with phyletic
894 coupling strength. Notes: *Architecture* CheA.X did not appear in the top 4% strongest
895 phyletic couplings and was excluded from the graph. cheb.F2, cher.F2, and mcp.48H
896 formed a cluster disconnected from the rest of the network, but all three coupled to
897 CheA.II at slightly lower strengths (top ~10%) (Dataset S4).

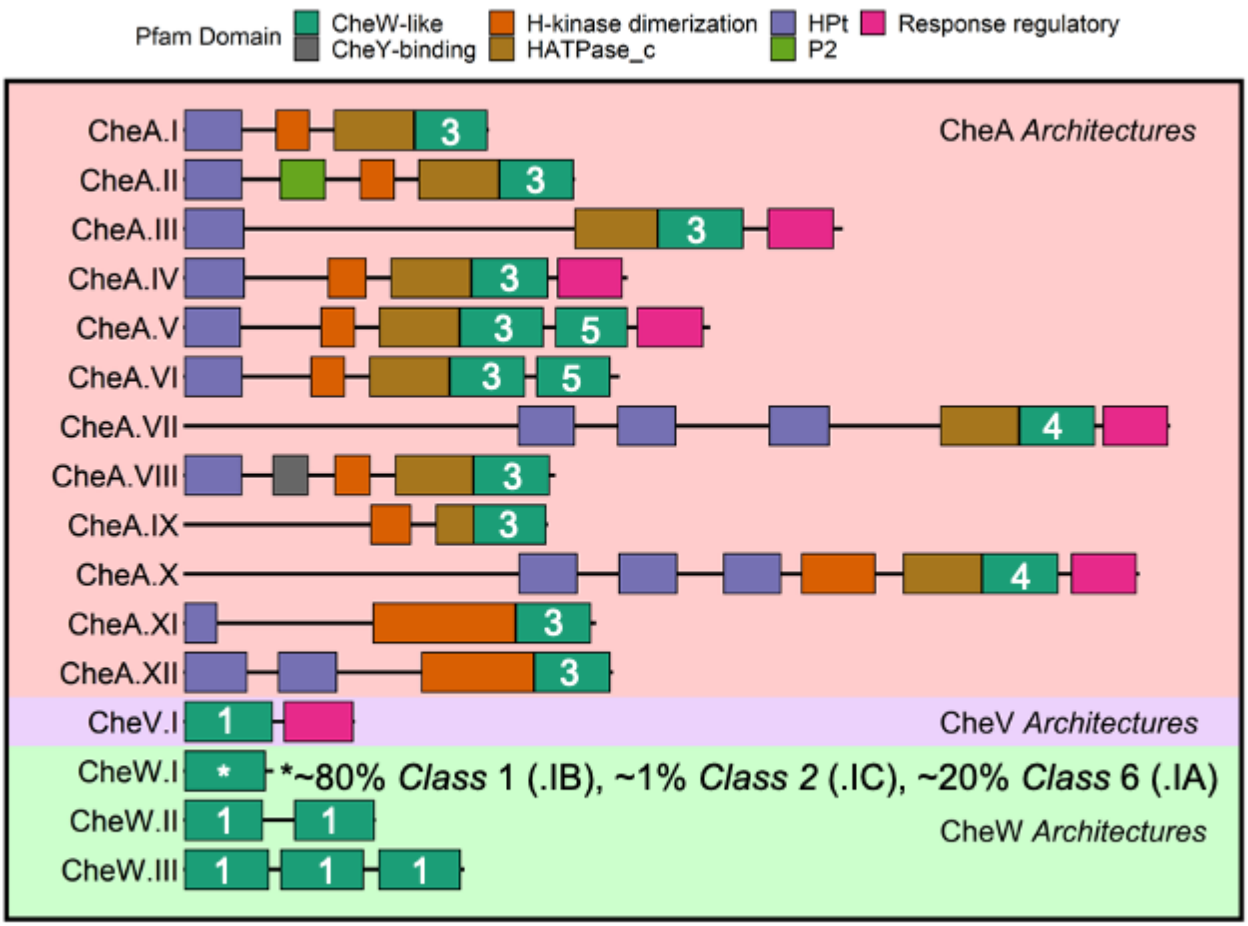


FIGURE 1. Major *Architectures* of proteins that contain CheW-like domains (from [36]). CheA-lineage *Architectures* are designated by a Roman numeral suffix in order of decreasing abundance. CheW-lineage *Architectures* are designated by a Roman numeral suffix indicating the number of CheW-like domains. The CheW.I *Architecture* includes sequences that belong to three distinct *Classes*, designated CheW.IA, CheW.IB, and CheW.IC. Classes of CheW-like domains are indicated by white numerals.

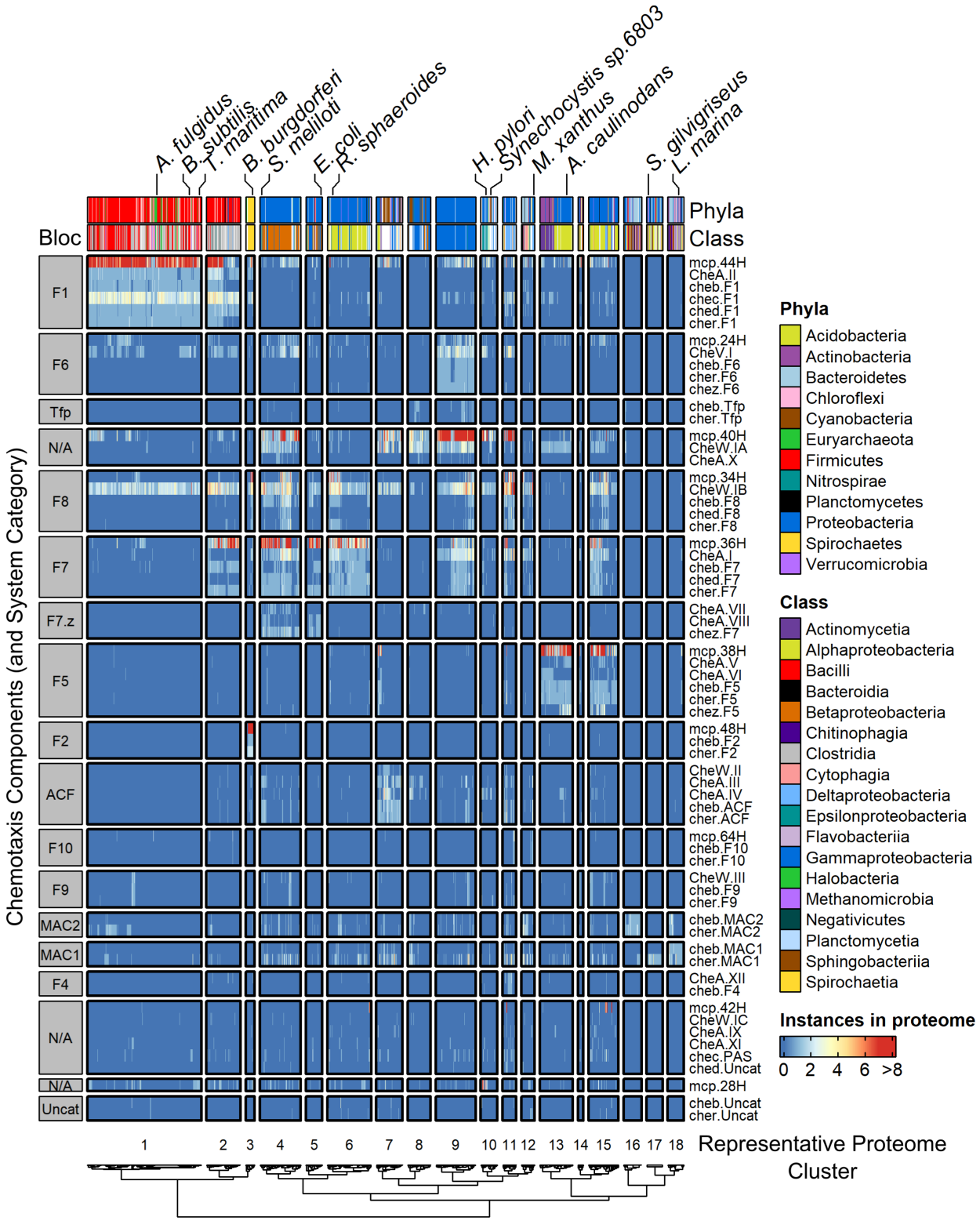


FIGURE 2. Co-occurrences of individual chemotaxis components in RP35 representative proteome set. Total occurrence counts were used. Components with <

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 shown in Figure S1.

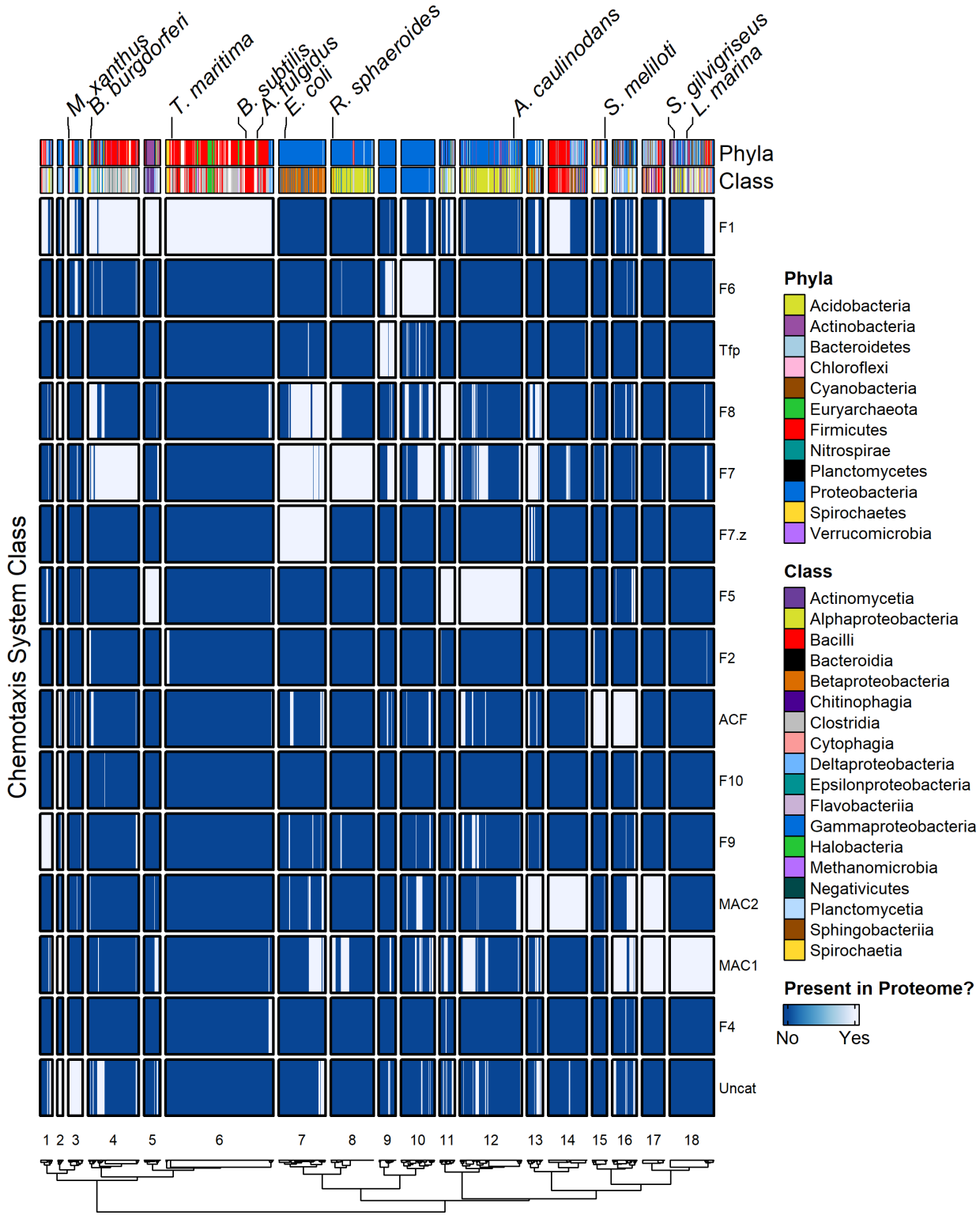
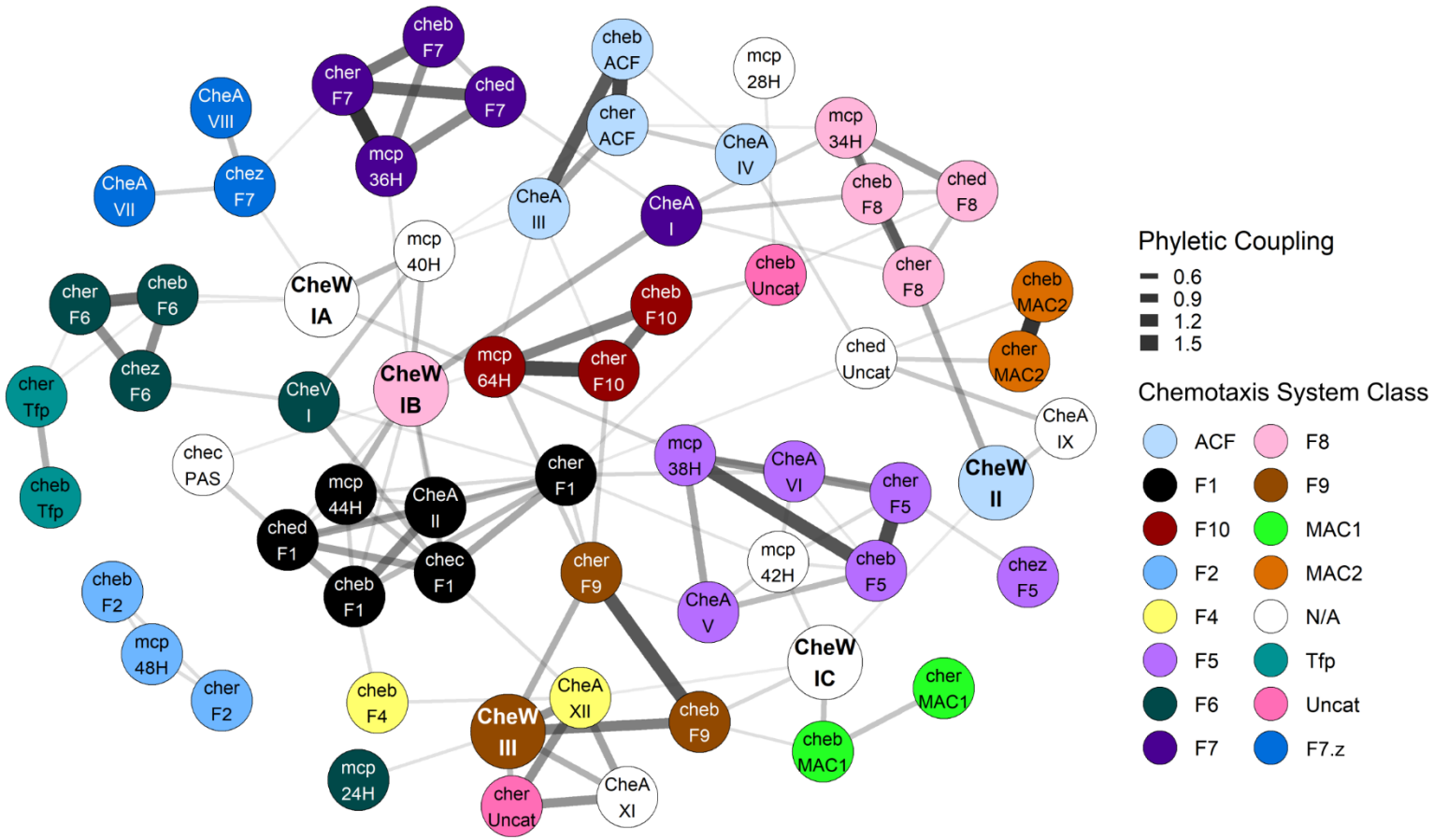


FIGURE 3. Simplified co-occurrence schematic of chemotaxis system categories in RP35 representative proteome set. A binary presence/absence scheme was used

919 for visualization. A chemotaxis system category was determined to be present in a
920 given proteome if at least one of the following components was detected of the
921 appropriate category: CheB/C/D/R/Z. CheA/V/W and MCP components were excluded
922 from the analysis, because some of these components function with more than one
923 chemotaxis system category. Notable organisms were labelled. Row order was
924 maintained for consistency with Figure 2, meaning categories were not clustered.
925 Results were split by dendrogram height into functional “blocs” by clustering proteomes
926 (columns). However, because the datasets upon which Figures 2 and 3 are based are
927 different, the resulting proteome clusters and cluster numbers are different than in
928 Figure 2.



929

930 **FIGURE 4. Network representation of inferred phyletic couplings between *Architectures* containing CheW-like**
931 **domains and remaining chemotaxis system components.** Co-occurrence data of chemotaxis components extracted
932 from the RP35 representative proteome set were converted to a binary phylogenetic profile matrix. Phyletic Direct
933 Coupling Analysis (PhyDCA) was used to quantify the favorability (correlation) of chemotaxis components co-occurring
934 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 (~4%) 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 4% 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).