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

3/2/22

Generalizable strategy to analyze domains in the context of parent protein architecture: A CheW case study

Luke R. Vass[1,#a], Katie M. Branscum[#b], 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: Domain variation with protein architecture

*Corresponding author

Email: bourret@med.unc.edu

Co-author email addresses:

Luke R. Vass: lukev3@vt.edu

Katie M. Branscum: katief@tribaldiagnostics.com

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

Data Availability Statement:

The data that support the findings of this study were derived from the following resource available in the public domain: Pfam, http://pfam.xfam.org.

Conflict of Interest Statement:

The authors declare no conflict of interest.

ABSTRACT

Domains are the three-dimensional building blocks of proteins. An individual domain can be found in a variety of protein architectures that perform unique functions and are subject to different evolutionary selective pressures. We describe an approach to evaluate the variability in amino acid sequences of a single domain across architectural contexts. The ability to distinguish the different evolutionary paths of one protein domain can help determine whether existing knowledge about a specific domain will apply to an uncharacterized protein. Such discrimination can lead to insights and hypotheses about function, as well as guide experimental priorities.

We developed and tested our approach on CheW-like domains (PF01584), which mediate protein/protein interactions and are difficult to compare experimentally. CheW-like domains occur in CheW scaffolding proteins, CheA kinases, and CheV proteins that regulate bacterial chemotaxis. We chose 16 protein *Architectures* that included 94% of all CheW-like domains found in nature. Because some *Architectures* had more than one CheW-like domain, CheW-like domains were divided into 21 distinct *Contexts*. The CheW-like domain sequences were closely related within most *Contexts*; however, one *Context* was subdivided into three *Types*. The resulting 23 sequence *Types* coalesced into five or six *Classes* of CheW-like domains, which we described in detail.

In addition, we created SimpLogo, an innovative method for visualizing amino acid composition across large sets of multiple sequence alignments of arbitrary length. SimpLogo offers substantial advantages over standard sequence logos for comparison and analysis of related protein sequences. The R package for SimpLogo is freely available.

KEYWORDS

protein domain, protein architecture, CheW, CheA, CheV, SimpLogo

1 | INTRODUCTION

Many proteins consist of a chain of discrete domains. Protein domains are typically compact, independently folding units. Originally, protein domains were defined based on protease resistance; however, they are now recognized by characteristic amino acid sequence features. Individual domains possess structures and functions that are subject to natural selection and thus are discrete units of evolution. Domains were assembled in many permutations during evolution to create different proteins. Therefore, domains can be considered modular “building blocks” of proteins. The Protein Family (Pfam) database currently contains almost 20,000 different domains.[1]

Identifying domains is useful because once nature has evolved the solution to a problem (e.g., a protein domain that performs a specific function), the solution is used repeatedly. Thus, domain identity often provides useful general predictions about the structure and function of domains in experimentally uncharacterized proteins, as well as providing an effective means by which to compare proteins of known function. The success of such predictions relies on the accuracy of domain descriptions, which are often considered to be monolithic. While useful in some respects, in actuality there can be substantial differences between domains of the same type that are found in different proteins. For example, response regulator proteins are formed by linking a receiver domain (Response_reg, PF00072) to a wide variety of output domains. At least three types of evidence demonstrate the differences between receiver domains in different types of response regulators: (i) phylogenetic analysis shows receiver domains co-evolve with their output domains,[2,3] (ii) receiver domain:output domain interaction surfaces and receiver domain dimerization surfaces vary as a function of binding

partner and whether a receiver domain exerts positive or negative regulation on its output domains,[4] and (iii) receiver domains support phosphorylation and dephosphorylation reaction kinetics that differ by many orders of magnitude depending on the timescale of the associated biological function.[5-8] Thus, receiver domains, though termed the same, differ depending upon the protein architecture in which they are found.

Variation between different instances of a particular domain raises multiple questions. If the structure, function, and/or other characteristics of a domain are experimentally characterized in one protein, then how well does that information translate into useful insights about the same domain in another protein? In other words, how does a specific protein domain vary when found in the contexts of different protein architectures? Genome sequencing has far outstripped experimental capacity but provides a rich source of relevant information. Can sophisticated methods of amino acid sequence analysis identify contexts (i.e., location within protein architectures) in which current knowledge about a specific protein domain applies? Conversely, given the enormous scope of uncharacterized proteins and protein architectures, can we identify contexts where current knowledge about a specific domain may not apply? Such a method would provide a way to inspire and focus future experimental investigations of protein domains in a manner that is most likely to be informative and may also lead to useful hypotheses and testable predictions. We describe here our development of a generally applicable approach to analyze the variation in a specific protein domain as a function of the contexts in which the domain is found, revealing insights at a substantially higher resolution than a typical imprecise Pfam domain assignment.

As a test case to develop, validate, and illustrate our method, we chose CheW-like domains (CheW, PF01584), which are part of the signal transduction pathways that control bacterial chemotaxis.[9] The known function of CheW-like domains is to act as scaffolds for protein/protein interactions.[10,11] The absences of an active site, substrate or ligand binding, or enzymatic activity make CheW-like domains challenging to study experimentally because no simple assay exists by which to compare the properties of different CheW-like domains. One might envision measuring binding constants, but it would be difficult to untangle the contributions of the CheW-like domains and their partner proteins. Furthermore, CheW-like domains are used to assemble very large multi-protein complexes that couple transmembrane chemoreceptors to cytoplasmic kinases, so the biological relevance of the binding constant between an isolated CheW-like domain and its receptor or kinase partner is not immediately obvious. Finally, evolutionary constraints on CheW function are such that sequences can diverge substantially and still be functionally interchangeable.[12] Thus, an objective method to classify CheW-like domains from different contexts would be particularly helpful to clarify the circumstances in which current knowledge is likely to be applicable and those which would likely benefit from further investigation.

CheW-like domains are found in three broad lineages of protein architectures: CheW scaffold proteins, which consist only of CheW-like domains; CheA kinases, which combine CheW-like domains with domains that provide kinase activity and phosphorylation sites; and CheV proteins, which combine CheW-like domains with receiver domains. Notably, the function of CheV proteins is poorly understood. A recent review of the diversity of bacterial chemotaxis systems asked, “Can efficient

computational approaches be developed to distinguish … CheW from other CheW domain-containing proteins?”.[13] Our results suggest that the answer is yes, and that not all CheW proteins are alike.

2 | MATERIALS AND METHODS

2.1 | Protein sequence database

The Pfam database (version 33, obtained May 2020) was used for the purposes of this study.[1] Any sequence found within the Representative Proteome 35 (RP35) to contain the CheW-like domain (PF01584) was extracted for further use, retaining only the CheW-like domain sequence.[14] “35” refers to the co-membership threshold used to control the granularity of the sequence space represented. An RP set best represents the known sequence diversity of a group of similar proteomes/organisms. To construct the RP35 data set, available protein sequences for completed proteomes within the UniProtKB database were grouped based on their UniRef50 cluster membership. UniRef, or UniProt Reference Clusters, are sets of proteins grouped together by sequence in such a way as to produce tight clusters of orthologs/inparalogs, maximizing sequence space while minimizing redundancy; 50 refers to the sequence identity threshold used for the clustering procedure.[15] We chose RP35 as the best compromise between information diversity and sequence space reduction. CheW-like domain sequences were grouped into *Architectures* based on the domain organization of the parent protein (Figure 2). *Architectures* capturing fewer than 40 sequences were excluded from the analysis as uninformative.

2.2 | Sequence analysis and initial clustering

CheW-like domain sequences were collectively aligned to the current CheW-like domain HMM (PF01584) using HMMER3 (to facilitate cross-architecture comparisons).[16] The removal of poorly aligned sequences or subsequences has been shown to be beneficial for large-scale, downstream sequence analysis and phylogenetic

profiling. To that end, we used trimAl (version 1.2) to remove spuriously aligned regions within our CheW-like domain sequences (using a column gap-threshold of 90%, assuming that the top 10% most “gappy” columns were mostly noise).[17] Entries with >50% gap-characters were also removed.

We then used the mat.dis() function from the bios2mds package in R (version 1.2.3) to generate symmetric, pairwise-dissimilarity matrices from our sequence alignment for further clustering (with zero values along the diagonal, representing a sequence’s dissimilarity to itself)[18] using four different evolutionary substitution models: GONNET, point accepted mutation 250 (PAM250; number of mutations per 100 amino acids), BLOcks SUbstitution Matrix 30 (BLOSUM30), and Jones-Taylor-Thornton (JTT).[19-22] Each is tailored towards higher diversity sequence groups (the average sequence identity reported in the full Pfam alignment for CheW-like domain PF01584 is only ~20%). The *k*-medoids clustering protocol was performed using the pam() function from the cluster package within R (version 2.1.2; *k*-medoid clustering is sometimes called Partitioning Around Medoids).[23] Our initial input parameter *k* was set to 21 (total number expected of distinct contexts for CheW-like domain; intentionally chosen as an overestimate). An optimal *k* value (~3-4) was also estimated using the average silhouette width method implemented in the pamk() function from the fpc package in R (version 2.2-9; *k*=3-10).[24] The similarity of the partitions generated using the four substitution models was quantified using both the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics.

ARI is a popular nonparametric statistic used in cluster validation to provide a quantitative measurement of the similarity between two partition schemes (regardless of

cluster nomenclature). It involves producing a general Rand Index (RI)[25] by calculating the ratio of the number of common element groupings between solutions using the following equation:

$$RI = \frac{a+b}{\binom{n}{2}} = \frac{correct\ same\ pairs + correct\ different\ pairs}{total\ number\ of\ pairs} \quad (1)$$

Here, the binomial coefficient $\binom{n}{2}$ represents the total number of unordered pairs in the set of *n* elements (equivalent to *n(n-1)/2*). In the equation, *a* represents the number of subset assignments common to both solutions, and *b* represents the number of assignments that are different. RI scores range from 0 (no overlap between clustering solutions) to 1 (perfect overlap). The more accurate Adjusted RI[26] essentially rescales the RI value to accommodate for the fact that random chance will produce some level of overlap between clusters using the following simplified scheme:

$$ARI = \frac{(RI - expected\ RI)}{(maximum\ RI - expected\ RI)} \quad (2)$$

This relationship ensures that for a completely random assignment of clusters and samples, the ARI score will equal zero.

Mutual information (MI) is an information theoretic metric quantifying the mutual dependence between two variables, revealing the amount of information obtainable for one variable through the observation of another.[27] It is essentially a measure of correlation and uncertainty; high MI suggests a large decrease in uncertainty about one random variable based on data from another random variable, whereas zero MI suggests that the two variables are independent. The measure of this uncertainty is called entropy. The MI between random variables X and Y is given by the simplified equation:

$$MI(X;Y) = H(X) - H(X|Y) \tag{3}$$

Here, $H(X)$ is the entropy of the single random variable X and $H(X|Y)$ is the conditional entropy of variable X conditional upon the knowledge of variable Y. A normalized variant of this (NMI) is often used to compare clustering performance.[28] The normalized form of MI is defined by the following equation:

$$NMI(X;Y) = \frac{H(X) + H(Y) - H(X|Y)}{\big(H(X) + H(Y)\big)/2} \tag{4}$$

NMI has the advantage of scaling the MI measurement within an upper bound of 1 (perfect correlation) and a lower bound of 0 (no correlation), making interpretation more intuitive. Two partitioning schemes with NMI scores close or equal to 1 are nearly identical.

2.3 | Multidimensional scaling analysis

Multidimensional scaling (MDS) analysis was performed using three different algorithms. Metric MDS methods focus on the “distances” in the matrix being analyzed, whereas non-metric MDS methods focus on the ranking in the matrix. Due to its simplicity and previous success when applied to protein sequences, we chose to use a metric MDS algorithm[18] to analyze sequence data from *k*-medoid clustering. For sequence subsampling, we used two MDS methods. The first was a implementation of the metric algorithm featured in the wcmdscale() function from the vegan package in R (version 2.5-7).[29] Given the large number of negligible negative eigenvalues generated using the metric implementation, we speculated that our initial pairwise distances may have been non-Euclidean in nature. Consequently, we tested a non-metric implementation of the MDS algorithm using the metaMDS() function, also from the

vegan package (using parameters: *trymax = 2000, noshare = F, maxit = 9999, sratmax = 0.99999999, k = 3, model = "local", autotransform = F, wascores = F, center = T, pc = T*) [29]. Resulting plots were visualized using the ggplot2 package in R (version 3.3.3).[30]

2.4 | Sequence subsampling strategy

Random sequence subsampling was performed on the various CheW-containing *Types* to facilitate dimensionality reduction analysis. Each architectural *Context* was treated as a separate homogeneous class, except for CheW.I, which was split into CheW.IA, CheW.IB and CheW.IC *Types*. CheW-like domains from parent *Architectures* featuring multiple domains were split into separate *Contexts* with appropriate suffixes added relative to their N- to C-terminal position in the parent protein. Two main approaches to stratified subsampling exist: randomly selecting a predefined number of observations from each defined subgroup; or randomly selecting a number of observations proportional to the original subgroup sizes. Given the wide disparity in the number of CheW-like domain sequences of each *Type*, we chose the former to facilitate computational efficiency. The mat.dis() function from the bios2mds R package was incorporated into an iterative sampling procedure.[18] Each subsampling round randomly selected 25 sequences from each *Type* (without replacement; smaller subsets were also tested with similar results). Sampling was repeated 10 times, generating a total of 10 replicate sequence sets for each of the four previously described substitution models. The function also calculated pairwise dissimilarity matrices for each replicate set for further use during the MDS analysis.

2.5 | Conservation analysis and SimpLogo creation

All CheW-like domain sequences were first aligned collectively to the conserved CheW-like Hidden Markov Model (HMM) profile (PF01584) obtained from the Pfam database. Separate sequence alignments were then segmented based on corresponding types of CheW-like domains. For each alignment, the aligned, standalone *Escherichia coli* CheW sequence was inserted at the top for use as a reference point. Columns introducing gaps (relative to the *E. coli* CheW sequence) were removed for simplicity. Position-wise count matrices and corresponding frequencies were generated for each alignment using the Biostrings and motifStack packages in R.[31,32] Physicochemically similar residue types were combined to reduce visual complexity of the final representation, creating nine amino acid groups (including gap characters): positively-charged side chains (His/Lys/Arg), negatively-charged side chains (Asp/Glu), small volume side chains (Ala/Gly), polar side chains (Gln/Asn/Ser/Thr), large hydrophobic side chains (Met/Val/Leu/Ile), gaps (-), aromatic side chains (Phe/Tyr/Trp), Pro, and Cys. Each amino acid category was assigned an "idealized" color and converted to tristimulus values in the CIE XYZ color space. XYZ space is linear and non-negative (as opposed to traditional RGB space), making it simple to mix colors. Final values were calculated by weighting (multiplying) the idealized colors of each amino acid group by their relative frequencies and summing the results for each position. Logos were visualized using simple geom_tile objects (from the ggplot2 suite), with colors and heights for each position based on the aforementioned blended tristimulus values and their overall gap character frequency, respectively. "Idealized" (i.e., color at 100% frequency) color strips for the top two most frequent residue types were added to each position (at the top and bottom of each tile)

to mitigate uncertainty between true qualitative colors and blended hues. Position-wise information content (in bits) was also estimated for each alignment again using motifStack and visualized with a line trace at the top.[32] To create traditional sequence logos (for comparison purposes), the Logolas R package was used with default settings to generate figures with similar dimensions.[33] An R package for SimpLogo was created and is available (along with an example and the sequence alignments used to generate Figure 5) in the Github repository https://www.github.com/clayfos/SimpLogo. We used Zenodo to assign a DOI to the repository: https://doi.org/10.5281/zenodo.6323714.

2.6 | Protein homology modeling and structural analysis

To visualize the relevant functional surfaces of the reference *E. coli* CheW-like domains CheA-P5 and standalone CheW, we modeled trimeric CheW:CheA-P5:receptor complexes (capturing both interface 1 and 2) based on the crystal structure of the *T. maritima* multimer featuring CheW, CheA-P5 fragment, and the chemoreceptor TM14 (PDB ID 4JPB).[34] We first generated homology models of *E. coli* CheA-P5 (residues 514-640; using the crystal structure of *T. maritima* CheA PDB ID 1B3Q[35] as a template) and CheW (residues 18-154; using the solution structure of *E. coli* CheW PDB ID 2HO9 as a template) with the Phyre2 web portal.[36] We then superimposed the top models against the corresponding *T. maritima* molecules in the crystal structure (PDB ID 4JPB[34]) using PyMOL (v.2.3.3).[37]

3 | RESULTS

3.1 | CheW-like domains are incorporated into dozens of distinct protein *Architectures* in nature

We first surveyed the known protein architectures into which CheW-like domains are incorporated, which provided initial sequence groupings for further investigation. We extracted the sequences of all proteins containing a CheW-like domain (PF01584; 8484 sequences containing 9094 unique CheW-like domains) from the Pfam database (version 33, accessed in May 2020)[1] using the Representative Proteome 35 (RP35) sequence set.[14] The use of RP35 provided an efficient and representative data set while substantially reducing the sequence space and complexity required for the analysis. CheW-like domain sequences from rare architectures (<40 member sequences) were excluded because we anticipated that interpretation would be too speculative. The remaining sample contained 8581 CheW-like domains (94% of original) in 8061 proteins (95% of original). Additional curation of sequences described in Materials & Methods (e.g., for excessive gap character) resulted in final analysis of 8528 sequences from CheW-like domains (94% of original).

We use the capitalized and italicized terms *Architecture*, *Context*, *Cluster*, *Group*, *Type*, and *Class* to indicate specific ways of categorizing sequences of CheW-like domains in this study. The relationships between the classification terms, as well as the flow of our sorting scheme, are outlined in Figure 1. We first categorized CheW-like domain sequences into three “lineages” based on the architecture and putative role(s) of their parent protein (Figure 2). Our nomenclature for *Architectures* containing CheW-like domains used the accepted lineage of the parent protein (e.g., CheA) as a prefix

and a Roman numeral as a suffix. CheA-lineage *Architectures* were numbered by decreasing membership size, starting with Roman numeral I. Additionally, *Architectures* containing two or more CheW-like domains (e.g., CheA.V from Figure 2) were split into their constituent CheW-like domains and analyzed as separate sequence *Contexts*. An additional suffix of an Arabic numeral identified the “position” of the CheW-like domain relative to the N-terminus and the other CheW-like domains in the parent protein, e.g., CheA.V.1 (*Contexts* are not labeled in Figure 2). The CheA-lineage was the most diverse, with a dozen major *Architectures*. In addition to CheW-like domains, most CheA *Architectures* also contained HATPase_c domains to bind ATP and catalyze phosphorylation, and histidine phosphotransferase (Hpt) domains to provide a phosphorylation site.

Although several CheV-lineage architectures (named for the canonical *Bacillus subtilis* CheV protein) were detected in the RP35 set, the only organization incorporating an appreciable number of sequences was the canonical CheV protein (termed the CheV.I *Architecture* for consistency). The CheV-lineage was the least abundant of the three architectural lineages.

CheW-lineage architectural variants constituted the largest collection of sequences, with over half of the entire data set attributable to canonical, standalone CheW proteins from which the *Architecture* gets its name (CheW.I). Additional rare (~1 to 3% the number of CheW.I) *Architectures* featured two or three CheW-like domains fused together (termed CheW.II and CheW.III respectively) with no other detectable domains present. Individual CheW-like domains of CheW.II and CheW.III *Architectures* were separated into distinct *Contexts* for further analysis, doubling or tripling the number

of sequences for the corresponding *Architecture* (i.e., CheW.II.1, for the first position CheW-like domain, CheW.II.2 for the second position CheW-like domain, etc.).

3.2 | CheW-like domains segregate into distinct sequence *Groups* as a function of parent protein *Architecture*

We expected that CheW-like domains extracted from a common architectural *Context* would more closely resemble each other rather than the same domains obtained from contrasting architectural variants. We based our assumption on existing knowledge of microbial response regulator receiver domains, which cluster phylogenetically predicated on the presence and/or type of an attached output domain.[2,3] However, the full dataset contained over 8500 CheW-like domain sequences, making analysis and interpretation challenging. Ideally, each CheW-like domain *Context* would behave as a homogeneous sequence set, and a basic stratified sampling strategy could be used to select a representative and more manageable population for study. While the assumption of homogeneity seemed reasonable (shared architecture typically implies common function), we wished to first gain insight into the underlying structure of the sequence space by performing unsupervised clustering on the full list of sequences. Unsupervised clustering would allow us to intelligently select the best representative sample from our data set. In essence, our approach would be a simplified multistage sampling scheme, first clustering the entire population to reveal major strata within each architectural *Context*, then randomly sampling from within these strata. This would ensure that all major sequence groups were represented. We chose to use a *k*-medoids approach for the initial clustering procedure to mitigate the difficulty of visualizing over 8500 data points. The *k*-medoids algorithm is a classical

clustering technique that partitions the data into *k* number of *Clusters*, attempting to minimize the distance between data points and the center of their relevant *Cluster*. Each *Cluster* is represented by a single central data point called a cluster medoid, which allowed us to visualize the *Clusters* using just the medoid sequences, rather than the entire data set.

Cluster analysis of protein sequences can be performed in a number of ways, though it most commonly involves using a multiple sequence alignment to calculate dissimilarity scores between all possible pairs of sequences and clustering the resulting matrix. We generated symmetric, pairwise-dissimilarity matrices from our sequence alignment for further clustering as described in Methods. Estimating protein sequence (dis)similarity requires the use of a substitution matrix to describe the rate or frequency at which a residue is substituted for another over time, simulating the process of evolution. Many types of substitution matrices exist. Because we were uncertain which would be most appropriate for our CheW-like domain data set, we tested four distinct substitution models as described in Methods. We calculated four distinct dissimilarity matrices for our sequence alignment using the four substitution models as inputs for the *k*-medoids cluster analysis. For convenience, we only show solutions derived from the GONNET substitution matrix in the main text, due to the documented ability of GONNET to handle sequences of greater evolutionary distance.

The *k*-medoid technique requires an initial target number of clusters (the *k*-value) as input. An optimal *k*-value can be estimated with various metrics, such as the average silhouette width (ASW; a measure of how similar a point is to its own cluster compared to other clusters), but choosing the “correct” number of clusters is challenging,

subjective and dependent on the overall goals of the analysis.[38] *A priori*, we chose *k* = 21 (the total initial number of distinct architectural *Contexts*, including multiple domains in the same protein, used to partition the CheW-like domain sequences) as our *k*-value. This was an intentional overestimation of the true number of subpopulations (optimal *k*-value estimation using the ASW suggested 3-4 clusters was ideal, data not shown). We then performed clustering as described in Methods. Pairwise dissimilarities between the medoids for each of the 21 identified *Clusters* were extracted and used as input for further analysis with a metric implementation of multidimensional scaling (MDS) as described in Methods. MDS is a multivariate dimensionality reduction and visualization technique that takes pairwise similarities/dissimilarities/distances for a data set and attempts to reconstruct a simplified map that maintains the distances between data points and is easily interpretable (i.e., in a lower *k*-dimensional space). In metric MDS (sometimes more generally referred to as Principal Coordinates Analysis), we attempt to visualize the (dis)similarity data along the dimensions that maximize the amount of variation (or information) within the data, similar to Principal Component Analysis.

A representative MDS plot describing just the 21 extracted medoids (based on the dissimilarities calculated using the GONNET model) is shown in Figure 3. Additional plots generated using dissimilarities calculated with the remaining substitution matrices are in Figure S1. Letters naming the *Clusters* for each solution were assigned based on decreasing number of sequences in the *Cluster*. There were three substantial *Groups* of *Clusters* of CheW-like domains (labeled 1, 3, or 4 in Figure 3), and another smaller, intermediate *Group* separating *Groups* 1 and 3 (labeled 2 in Figure 3, more apparent when considering only the first dimension). The partitions were highly similar between

the four clustering solutions (Figures 3 and S1), with pairwise Adjusted Rand Index (ARI) scores ranging from 0.6-0.7 and Normalized Mutual Information (NMI) scores of 0.7-0.8 calculated as described in Methods. Consistently high ARI and NMI values across solutions derived using four different substitution matrices suggested a relatively stable underlying data structure.

We next examined the composition of the individual *Clusters* to determine if the original architectural *Contexts* of CheW-like domains separated into multiple stable strata. Although each medoid was representative of the dissimilarity data, a *Cluster* may contain multiple domain *Contexts*. Understanding the composition of each *Cluster* would allow us to properly organize sequences for analysis on the more complete sequence set, rather than on just the representative medoids, ensuring that no distinct *Contexts* would be underrepresented. Table 1 assigns *Clusters* to the four main *Groups* indicated in Figure 3 and contains a detailed breakdown of the architectural *Contexts* of the sequences in each *Cluster*.

Multiple patterns are apparent in Table 1. First, sequences form *Clusters* and *Clusters* form *Groups* primarily based on parent protein lineage. Most or all sequences within a *Cluster* are from either CheA proteins or CheW/CheV proteins, not both. Most CheW-like domains from CheA proteins are in *Group* 1 and most CheW-like domains from CheW and CheV proteins are in *Group* 4. Closer inspection reveals additional structure in the data:

- Sequences of CheW-like domains from most CheA-derived *Contexts* were found in *Group* 1.

- The “second” position CheW-like domains from *Architectures* CheA.V and CheA.VI formed their own distinct sequence *Group* 2, made up of *Cluster* P.
- CheW-like domains from *Architectures* CheA.VII and CheA.X comprised the bulk of *Cluster* R, the smallest *Cluster* in *Group* 1.
- About 80% of domains from CheW.I *Architectures* and greater than 90% of domains from CheW.II and CheW.III *Architectures* were in *Group* 4.
- About 20% of domains from CheW.I *Architectures* and 5% of domains from CheW.II and CheW.III *Architectures* were in *Group* 3, possibly distinct from the majority of CheW-lineage domains.
- About 1% of domains from CheW.I *Architectures* and 2% of domains from CheW.II and CheW.III *Architectures* were in *Group* 1, suggesting similarity to the CheA-lineage.
- About 99% of CheW-like domains from CheV.I *Architectures* were in *Group* 4, suggesting that the presence of the attached receiver domain has little effect on the primary sequence of the CheW-like domain in CheV proteins.
- Similar to CheW proteins, about 1% of CheW-like domains in CheV.1 *Architectures* were in *Group* 1, suggesting similarity to the CheA-lineage.

Based on the analysis of Table 1, we assumed for further processing that each architectural *Context* could be reasonably treated as a single homogeneous entity, except for CheW.I. We split CheW.I into three distinct *Types* designated CheW.IA (sequence *Group* 3), CheW.IB (sequence *Group* 4), and CheW.IC (sequence *Group* 1). Remaining outlier sequences in each architectural *Context* (typically a small number of sequences in each *Context* assigned to substantially different *Clusters* than the

majority) were excluded from the analysis as noise. We believe that our evaluation produced an accurate summary of the underlying data structure, though we acknowledge the somewhat subjective nature of cluster analysis and that alternative perspectives may offer equally valid insights. We found the visualization in Figure 3 to be limited in interpretability when searching for biologically relevant trends. To address this shortcoming, we next performed MDS analysis on a larger, randomly sampled subset of our sequence data, now organized into 23 *Types* (due to CheW.I sub-designations) instead of 21 *Contexts*.

3.3 | Random stratified subsampling facilitates efficient multi-dimensional scaling analysis on large protein sequence sets and reveals *Classes* of CheW-like domains

One technical disadvantage of the MDS technique is its substantial computational cost. The time complexity of standard metric MDS is proportional to the square of the input size, which indicates significant temporal and technical requirements when analyzing a high-dimensional data set with thousands of data points.[39] Thus, the MDS technique is impractical when analyzing a data set such as our original ~8500 x 8500 dissimilarity matrix. One popular approach to dealing with large, unwieldy data is random subsampling, called stratified sampling when working with distinct groups or strata within the original population. For this strategy to be effective, adequate representation of each stratum is essential. Our initial cluster analysis (Figure 3 and Table 1) revealed how to organize sequences of CheW-like domains into 23 *Types* that should ensure all major subpopulations receive appropriate representation in such a sampling scheme.

We randomly chose 25 sequences from each of the 23 previously described *Types* to analyze with metric MDS as described in Methods. We repeated this stratified random sampling procedure for 10 iterations (575 distinct sequences per iteration) to assess the stability of the various randomized MDS solutions. We also included the CheW-like domains of 18 well-studied reference proteins as an internal constant for each iteration (data not shown).

Figure 4 shows the top three dimensions of one representative solution derived from the GONNET substitution matrix. The remaining nine iterations, as well as solutions calculated using randomly sampled subsets for the other three previously described substitution matrices, are in Figures S2-S13. The solution shown in Figure 4 revealed similar relationships to those found in Figure 3/Table 1, which we now designate as *Classes* (sub-sampling MDS) rather than *Groups* (*k*-medoid MDS). On the right area of the plot (Left and Center Panels), most CheW-lineage and CheV-lineage sequence *Types* (*Class* 1) clustered in a densely packed region. A large agglomeration of sequences from CheA-lineage *Types* (*Class* 3) was near the left side of the plot area (Left and Center panels). The CheA.V.2 and CheA.VI.2 *Types* (*Class* 5) were markedly different from the other CheA-lineage *Types*, consistent with previous assignment to *Cluster* P as the sole member of *Group* 2 in Figure 3/Table 1. Additional differentiation within the CheA-derived *Types* could be observed more effectively in the Center and Right Panels (incorporating Dimension 3), with the CheA.VII and CheA.X *Types* (*Class* 4) separating from the main body. This supports the clustering results presented in Table 1 and Figure 3, with the majority of CheA.VII and CheA.X sequence *Types* found in the somewhat independent *Cluster* R. Sequences belonging to the rare CheW.IC

Type (*Class* 2) were spread across Figure 4, alongside *Class* 1 CheW/CheV sequences, *Class* 3 CheA sequences, and in between (Left and Center Panels). In Figure 3/Table 1, a large subset of CheW-lineage sequences (*Group* 3) separated from most CheW- and CheV-lineage sequences (*Group* 4). The *Group* 3 sequences are termed *Type* CheW.IA (*Class* 6) in the analysis of Figure 4 and surprisingly did not separate from the main CheW/CheV *Types* (*Class* 1).

Other forms of MDS may offer additional advantages under the appropriate circumstances. Examination of our randomly sampled MDS solutions indicated that a significant proportion of the total eigenvalues were negative (~15.4-15.7%). Though the individual negative eigenvalues were miniscule compared to the positive entries, it was possible that our initial dissimilarities were non-Euclidean in nature, and another method such as non-metric MDS might offer additional insight. As a test, we attempted to derive further solutions using a non-metric implementation of MDS as described in Methods. These can be found in Figures S14-S25 in Supporting Information. We encountered significant challenges related to solution convergence, likely due to the still relatively large and complex data set (despite our subsampling strategy). However, the non-metric MDS solutions were largely reminiscent of the basic layouts generated using the metric algorithm, even when sampling smaller subsets of sequences. Thus, we concluded that the data structures were likely consistent irrespective of the MDS algorithm. Given the clear sequence heterogeneity within the CheW-like domains, we next sought to determine factors within the domain itself that distinguished the various *Types*.

3.4 | SimpLogo: A simplified method for visualizing large numbers of distinct protein sequences aligned over many positions

Arguably the most popular method for comparing sequence groups and their alignments is the sequence logo. In this context, a logo is a graphical representation depicting sequence conservation by stacking relevant characters, such as single-letter codes for amino acids or nucleotides, at each position in an alignment.[40] The relative heights of the characters are proportional to their frequency at the given position. Additionally, the overall height of a stack is often scaled based on the information content stored at that position within the alignment. Typical applications for amino acid sequence logos include visualizing functional motifs or binding sites within a protein family/group. However, the volume of information and the graphical complexity of a logo often makes it challenging to interpret when examining regions longer than two dozen positions. Sequence logos are less practical when comparing a large group of separate alignments (>12) with many positions (>50). A more ideal and complementary visualization technique would allow us to first evaluate the macroscopic characteristics of each sequence group. Subtle differences could then be examined in greater detail for specific regions using a more traditional approach like a true sequence logo, or even the original sequence alignment. Here, we describe a visualization format (colloquially referred to as SimpLogo) designed to compare the CheW-like domain sequences from the numerous *Types* described above. We sought to reduce the 23 individual sequence alignments (*Types*) into simplified, pseudo-one-dimensional representations. We accomplished this by weighting and blending qualitative colors assigned to specific amino acid residue types (grouped by physicochemical properties into eight classes)

based on their frequency at a given position within the alignments. We also modified the overall height based on its proportion of gap-character (details in Methods). This produced a single row of colored “boxes” for each sequence *Type*, reducing a complex problem to basic pattern recognition. To differentiate between secondary colors and a near-equal mixing of primary colors, colored strips were included on the top and bottom of each box, indicating first and second most frequent residue types. The simplified scheme allowed for the simultaneous comparison of dozens of distinct sequence families with lengths over many positions (>100) and encoding substantially more information about intermediate frequencies than simple consensus sequences. Additionally, we created traditional sequence logos (using the Logolas package in R) for the same alignments using similar image dimensions for comparison (see Figure S26). Highly conserved/informative positions are identifiable in the traditional logo, but the overall lack of space and reduced character sizes often make interpretation of specific positions challenging, especially for the lower frequency/less conserved positions with smaller stack heights. The patterns of gap characters (indicated by dashes) are also more challenging to interpret when compared to the SimpLogo representation in Figure 5. One could imagine how difficult interpretation would become with double the number of sequence groups or sequence lengths using traditional methods. Therefore, we envision the new SimpLogo visualization technique being used not as a replacement for traditional sequence logos, but as a supplementary technique for large scale exploratory analyses.

3.5 | *Classes* of CheW-like domains display characteristic gap patterns

The SimpLogo representation in Figure 5 revealed multiple trends. Overall, the relationships displayed in Figure 5 mimicked the trends seen in Figure 4, with most of the CheA-lineage *Types* clustered tightly (*Class* 3), the exceptions being the CheA.VII/CheA.X (*Class* 4) and CheA.V.2/CheA.VI.2 (*Class* 5) *Types*. Most CheW-lineage *Types* (*Classes* 1 and 6) and the CheV-lineage *Type* (*Class* 1) appeared similar, except for the CheW.IC *Type* (*Class* 2), which was intermediate in character between CheA and CheW *Types*. The gap character of the three canonical parent protein lineages likely played an important role in differentiating the sequence *Types*, particularly at positions 45-56 and 120-131. Here, we provide observations related to the corresponding gap patterns and offer possible functional implications. We assumed, for the sake of speculation, that the analogous interaction surfaces and behaviors are largely maintained throughout the various CheW-like domain topologies found in nature. All positions described are relative to *E. coli* CheW unless otherwise specified. Crystal structures of the CheW-like domains from well-studied *E.* coli CheW or CheA proteins are not available. To provide a structural reference for our observations, we constructed simple homology models of *E. coli* standalone CheW (residues 18-154) and the CheW-like domain (termed P5) of CheA (residues 514-640) using the Phyre2 web service.[36] Various structural alignments were performed against an existing crystal structure featuring CheW, CheA P5 domain, and the chemoreceptor TM14 from *Thermatoga maritima* (PDB ID 4JPB)[34] to provide insight into the functional consequences of the characteristics described in the following paragraphs. Table S1 contains a detailed list of relevant putative residue functions for convenient cross-referencing, based on various literature sources and our homology models.

At positions 45, 55, and 56, the CheA-lineage *Types* generally exhibited sequence gaps, consistent with the observation that sequences of CheW-like domains from CheA proteins typically appeared shorter overall than from single domain CheW proteins when using the standard Pfam model (corroborated in manual CheW-like domain sequence alignments, data not shown). Experimental binding data suggest that positions 45-56 are involved in the interaction between the standalone CheW and CheA, at least in *E. coli.*[41,42] The CheW-like domain topology resembles that of a SRC Homology 3 (SH3) domain and is made up of two separate, pseudo-symmetric subdomains (1 and 2).[43] Subdomain 1 of the CheW-like P5 domain in *E. coli* CheA (approximately residues 514-531 and 596-640) interacts with subdomain 2 of standalone *E. coli* CheW (approximately residues 36-113; termed interface 1), primarily through the β3-β4-β5 regions of each individual subdomain (approximately residues 608-633 on P5 and 44-67 on CheW, based on homology modeling).[34] However, the complex signaling array/lattice formed by chemoreceptor:CheW:CheA oligomers is facilitated by the formation of a secondary interface between subdomain 1 of standalone CheW and subdomain 2 of P5 (approximately residues 114-134 on CheW and 539-556 on P5, based on homology modeling; termed interface 2), an interaction that contributes substantially to the cooperativity of the signaling array.[44] Positions 45-56 did not overlap with known chemoreceptor binding surfaces on CheA or CheW, implying that differences in gap character seen in Figure 5 likely affect the primary and/or secondary associations between the CheW-like domains of CheA- and CheW-lineage *Architectures* and not direct interactions with the chemoreceptor. Because subdomain 2

serves distinct purposes in CheA and CheW, differences in the corresponding functional residues seemed reasonable.

There were two exceptions to the described dichotomy between CheA-lineage and CheW-/CheV-lineage *Types* regarding gaps at positions 45, 55, and 56. First, *Types* CheA.V.1 and CheA.VI.1 exhibited only a partial gap pattern at position 45. Recall that *Types* CheA.V.I and CheA.VI.1 fell into *Cluster* H (Figure 3/Table 1). Although ultimately considered part of *Group* 1, *Cluster* H showed significant deviation from the rest of the CheA-lineage *Types*, with increased resemblance to *Group* 2 (*Cluster* P, made up of *Types* CheA.V.2 and CheA.VI.2 from the same *Architectures*) in Figure 3. However, *Cluster P* exhibited gaps at position 45 (Figure 4). Second, the gap patterns at positions 45, 55, and 56 exhibited by *Type* CheW.IC were more reminiscent of the CheA-lineage sequences than the rest of the CheW-lineage *Types*. This is consistent with observations that (i) the CheW.IC *Type* at least partially resembled CheW-like domains from CheA-lineages (both belonged to *Group* 1 in Table 1) and (ii) as *Class* 2 spanned the sequence “gap” between CheA (*Class* 3) and CheW (*Class* 1) regions of Figure 4.

CheA-lineage *Types* (*Classes* 3-5) exhibited stretches of gaps at positions 120-131, which corresponds to subdomain 1 in CheW-like domains. Therefore, sequence differences here could also affect the association between CheA- and CheW-/CheV-lineage proteins. CheV- and CheW-lineage *Classes* 1 and 6 shared the gap pattern near position 120 with CheA-lineage *Classes* 3-5 but lacked the stretches of gaps in subdomain 1 near positions 124 and 130. In contrast, CheW.IC (*Class* 6) again more

closely resembled the gap profile seen in the CheA-lineage *Classes* 3-5 at positions 120-131.

Type CheA.V.2 and CheA.VI.2 domains (*Class* 5) displayed a slightly modified gap profile at positions 122-130 and exhibited a more diffuse level of gap character at positions just upstream of the region (indicated by subtly “shrunken” boxes stretching back to residue position 110). Although the role(s) played by positions 109-110 in array formation is unclear, structural and biochemical data suggest that residue 108 is directly involved in chemoreceptor binding.[45] In the dual CheW-like domain CheA.V and CheA.VI *Architectures*, the “first” position CheW-like domains (*Types* CheA.V.1 and CheA.VI.1) did not exhibit the same deviation in gap character at residues 110-119 as *Class* 5, suggesting that the “second” position CheW-like domains may bind chemoreceptors that are distinct from those of their N-terminal counterparts, or perhaps bind entirely different partners through this region.

3.6 | *Classes* of CheW-like domains display characteristic residue identities

Careful inspection of Figure 5 also revealed several characteristic deviations in residue frequencies at key positions, many of which involved *Class* 5 domains:

The entire region between positions 108-134 was filled with conserved charged and polar residues that are likely involved in intermolecular/interdomain interactions in CheA-lineage *Types* but that were absent in the other *Types*. At position 110, several of the CheA-lineage *Types* encoded a negatively charged (red) residue (particularly in CheA.II/III/IV/V.1/VI.1/VII/X), whereas the position was less conserved in the CheV- and CheW-lineage *Classes* 1 and 6. The gap character was increased at position 110 in *Class* 5. Based on our homology models and alignment against another crystal

structure of dimeric *T. maritima* CheA (residues 290-671 relative to the *T. maritima* protein; PDB ID 1B3Q), residue 110 likely interacts with the helical dimerization domain of the CheA kinase, possibly explaining the change in conservation in CheA-related *Types* compared to the other lineages. Similarly, position 113 was strongly hydrophobic (green), position 114 was strongly positively charged (blue), and position 115 was polar/Pro (yellow/white) in almost all the CheA-lineage *Types* except *Class* 5, but not in the CheV- and CheW-lineage *Classes* 1 and 6. These positions are likely related to formation of the interface between the CheW-like domains of CheA and CheW.

Class 5 domains exhibited distinct patterns of charged amino acids. Position 38 was positively charged (blue) in *Class* 5 but negatively charged (red) in all other *Classes* except 4. *Class* 5 also exhibited a pair of conserved, negatively charged (red) residues at positions 40-41, a unique trait not exhibited by any other *Class*. In *E. coli,* the region equivalent to position 40 is involved in chemoreceptor binding,[41,42,45] again suggesting different sensor/partner specificity for *Class* 5 domains.

At position 49, all four of the CheW-like domains in the CheA.V and CheA.VI *Architectures* shared a conserved negatively charged (red) residue. The remaining CheA-lineage *Types* exhibited substantially lower conservation, leading to brownish coloration, whereas the CheV- and CheW-lineage *Types* primarily exhibited a conserved Pro residue (white). Equivalent residues in *E. coli* are involved in the association of CheW-like domains in CheA and CheW, based on the homology models.

Class 4 *Types* (CheA.VII and CheA.X) did not appear obviously different than *Class* 3 *Types* (most CheA *Types*) in Figure 5 except for Pro at position 87. The only *Type* that consistently stood out in *Classes* 3 and 4 was *Type* CheA.VIII. At position 27,

most *Types* exhibited a low level of conservation, leading to a reddish-brown color. However, the CheA.VIII *Type* featured almost exclusively negatively charged (red) residue types. Closer examination of the traditional sequence logo and alignment for the CheA.VIII *Type* revealed a nearly invariant Glu. In our *E. coli* CheW homology model, the negatively charged Glu side chain of position 27 formed a hydrogen bond with the hydroxyl group of Tyr29 (brown) near the site of interface 2, which is essential for higher-order oligomer formation.

At position 41, there was an increased conservation of positively charged (blue) residues in many of the *Types* from CheA- and CheW-lineages, including CheA.V.1, CheA.XI, CheA.XII, CheA.VII, CheA.X, and to a lesser extent CheW.IB, CheW.III.1 and CheW.III.3. However, *Class* 5 domains exhibited an increase in negative charge (red) at the same position. *Type* CheA.VIII featured a highly conserved Gln (yellow) at position 41 followed by a similarly conserved Pro (white). The Gln Pro pair was not found in other *Types*. Experimental data and our homology models suggested that position 41 is involved in both CheA/CheW association and chemoreceptor binding.[42]

We did not observe any positions where *Class* 6 (CheW.IA), representing ~20% of CheW-lineage domains, differed from *Class* 1 (most CheW domains and CheV domains). However, we did observe some variation within *Class* 1:

- Position 56 exhibited gaps in most CheA-lineage *Types* and a strong hydrophobic (green) character in most CheW-lineage *Types* but exhibited increased negative charge (red) in CheA.VIII, CheV.I and CheW.III.3, and a positively charged ([blue) residue in CheW.IB. *In E. coli* CheW, experimental data indicate that position 56 is involved in the formation of interface 1, forming a

hydrogen bond with Tyr81. Our homology models suggested that the region is likely substantially altered in CheA-lineage structures, as indicated by the consistent gaps at positions 55-56,[41] with the exception of CheW-like domains from *Type* CheA.VIII (which includes the P5 subdomain of *E. coli* CheA). The deviations seen at position 56 were likely a result (or cause) of differences in primary CheW/CheA association interactions between various combinations of CheA and CheW *Types*.

- At position 32, most *Types* exhibited Pro (white), except *Types* CheW.IB, CheW.II.1, CheW.II.2, and CheV.I, which featured increased negatively charged (red) residues in the first three, and almost exclusively Asn (yellow) in the latter. Position 32 is proximal to the chemoreceptor binding interface (according to our homology model for the CheW-like domain of *E. coli* CheA) and is likely involved in the positioning and/or formation of corresponding intermolecular interactions. The high conservation of a Pro residue in most *Types* suggested an important structural feature.
- Position 74 featured a seemingly buried hydrophobic (green) residue in most CheW-like domains, with the exceptions of *Types* CheW.IB, CheW.II.1 and CheW.III.3, which instead displayed positively charged (blue) residues (primarily Lys). The entire region between positions 72-74 was significantly more positively charged in the CheW-lineage *Types*, with the possible exception of CheW.IA. The implications of this unique characteristic are unknown, because the positions are not known to be involved in receptor binding or interface formation.

- At positions 86-87 (a critical residue pair involved in chemoreceptor binding in *E. coli* CheW[41,45]), most of the CheW-lineage *Types* exhibited polar (yellow) and positively charged (blue) residues, except for CheW.II.1 and CheW.IC, which featured negatively charged/polar (red/yellow) and small-volume/hydrophobic (dark grey/green) residue pairs, respectively.

Finally, CheA.VI.1, CheW.II.2 and CheW.III.2 were the only *Types* with a strongly conserved Cys (cyan) residue, found at positions 24, 50-51, or 135 respectively. To the best of our knowledge, no naturally occurring disulfide bond formation has been reported involving a CheW-like domain, though conservation implied that the Cys residues were functionally or structurally significant in some way. The role of position 24 is currently unclear. Positions 50 and 135 are thought to be involved in the association of CheA- and CheW-lineage domains.

4 | DISCUSSION

4.1 | CheW-like domains sort into six *Classes*

We started with three lineages and 16 protein *Architectures* (Figure 2) that contained 94% of CheW-like domains. Due to multiple domains in four *Architectures*, there were a total of 21 *Contexts* for CheW-like domains. Cluster analysis (Figure 3) revealed that the largest *Context* (single domain CheW proteins) split into three *Types* (Table 1). Random stratified subsampling (Figure 4) showed that the 23 *Types* formed five or six *Classes* of CheW-like domains, a conclusion that was supported by visual inspection of SimpLogos (Figure 5):

Class 1 contained seven *Types* of CheW-like domains (CheV.I, CheW.IB, CheWII.1, CheWII.2, CheWIII.1, CheWIII.2, CheWIII.3) that comprised CheV proteins and most CheW proteins. CheW-like domains in CheW proteins were generally similar, regardless of whether the protein had one, two, or three CheW-like domains. CheW-like domains in CheV proteins were similar to CheW-like domains in CheW proteins.

Class 2 contained the ~1% of CheW-like domains in CheW proteins (*Type* CheW.IC) that shared similarity with CheW-like domains in CheA proteins (Table 1, Figure 5). In MDS analyses, members of *Class* 2 landed in the CheW region, the main CheA region, or in between (see Figures 4 and Figures S2-S25).

Class 3 contained 10 *Types* of CheW-like domains (CheA.I, CheA.II, CheA.III, CheA.IV, CheA.V.1, CheA.VI.1, CheA.VIII, CheA.IX, CheA.XI, CheA.XII) that encompassed almost all CheA proteins. CheW-like domains in most CheA proteins with one CheW-like domain were similar, even though there were 10 different major *Architectures* for such proteins, including one (CheA.IX) that lacked a phosphorylation

site (Hpt domain) and two (CheA.XI, CheA.XII) that lacked ATP binding sites (HATPase_c domain). Such “split” CheA *Architectures* are not degenerate, but can work together to reconstitute a functional unit.[46] N-terminal CheW-like domains in CheA *Architectures* with two CheW-like domains (CheA.V.1, CheA.VI.1) were similar to CheW-like domains in CheA *Architectures* with a single CheW-like domain.

Class 4 was made up of two *Types* of CheW-like domains (CheA.VII & CheA.X) from CheA proteins with three Hpt domains and a receiver domain. *Class* 4 CheW-like domains consistently diverged from other CheW-like domains in CheA proteins with a single CheW-like domain (Figure 4).

Class 5 consisted of C-terminal CheW-like domains in CheA proteins with two CheW-like domains (CheA.V.2 & CheA.VI.2). *Class* 5 CheW-like domains were different from all other CheW-like domains in CheA proteins (*Classes* 3 and 4), CheW proteins (*Classes* 1 and 6), and CheV proteins (*Class* 1) (Figure 4).

Visualization of *Class* 6 depended on the method used. Approximately 20% of CheW-like domains in CheW proteins (*Type* CheW.IA) appeared to be different from the rest by *k*-medoid analysis (see Figure 3/Table 1). *Class* 6 appeared indistinguishable from *Class* 1 in the metric MDS analyses of random stratified subsampling in Figure 4 and Figures S2-S13, and the differences were subtle in the SimpLogo of Figure 5. *Class* 6 showed some separation from *Class* 1 in the non-metric MDS analyses of random stratified subsampling in Figures S14-S25.

4.2 | Implications of CheW-like domain *Classes*

The existence of *Class* 1, which encompasses CheV and most CheW proteins, has multiple implications. First, evidence exists that CheV proteins can function as

scaffold proteins similarly to CheW proteins; however, more detailed information about the functions and mechanisms of CheV is sparse.[11,47,48] It is tempting to extrapolate what is known about response regulator proteins (composed of N-terminal receiver and C-terminal output domains) to formulate hypotheses about CheV (composed of C-terminal receiver and N-terminal CheW-like domains). However, our observation that CheW-like domains in CheV proteins were similar to those in CheW proteins constrains such hypotheses by suggesting that any interactions between the receiver and CheW-like domains in CheV proteins are not sufficiently specific to be subject to selective pressure for evolutionary optimization. There are at least three categories of hypotheses about CheV inspired by response regulators:

- The receiver domain might alter the conformation of or occlude access to the CheW-like domain, depending on the phosphorylation state of the receiver. Alexander *et al.* suggested that such phosphorylation-mediated changes could explain the role of CheV in adaptation by changing the conformation of the CheW-like domain to alter the degree of coupling between receptor and kinase function.[47] Our findings make this otherwise reasonable hypothesis now seem unlikely. In *Bacillus subtilis* CheV, the CheW-like domain inhibits phosphotransfer from CheA-P to CheV and stabilizes CheV-P from autodephosphorylation.[49] Our findings suggest that these effects are more likely the result of steric effects than conformational coupling between the two domains of CheV. The determination of CheV structures in the presence and absence of the stable phosphoryl group analog BeF_3^- would be highly informative.

- The phosphorylation site in the receiver domain of CheV might act as a phosphate sink, while the CheW-like domain localizes CheV appropriately.[47,48] Our results are consistent with such a hypothesis because no interaction between the two domains of CheV is required that might lead to divergence of CheW-like domains in CheW and CheV proteins. In a phosphate sink mechanism, when environmental circumstances change and phosphoryl groups stop flowing into the system from ATP, phosphoryl groups flow backwards from the CheY response regulator to CheA to the sink and are lost to hydrolysis. Localizing the sink near CheA would enhance efficiency. We note that another possible role of CheV is as a reservoir that retains phosphoryl groups in a buffering capacity.
- In a more dynamic role, phosphorylation-mediated dimerization of the receiver domain of CheV might control the ability of the CheW-like domain to bind with receptor and/or kinase partners. To the best of our knowledge, the multimeric state of CheV and any effect of phosphorylation on multimeric state have not been investigated.

Second, five (CheA.III, CheA.IV, CheA.V, CheA.VII, CheA.X) of the 12 CheA *Architectures* in Figure 2 contained a receiver domain immediately C-terminal to a CheW-like domain, which is reminiscent of CheV proteins. However, the fact that CheW-like domains in CheV and CheA proteins were in different *Classes* strongly suggests that adjacent CheW-like and receiver domains in CheA proteins do not act as CheV analogs. Consistent with this prediction, the adjacent CheW-like domain did not

affect autodephosphorylation of the receiver domain from *Azorhizobium caulinodans* CheA,[50] in contrast to *B. subtilis* CheV.[49]

Third, as far as we know, obvious questions about multi-domain CheW proteins have not been explored experimentally. It is not known if all domains of multi-domain CheW proteins are required for function, or if there are differences in functional roles between the domains. The fine structure of the data in Table 1 (specifically the number of sequences in a *Cluster* varied substantially as a function of *Context* within CheW.II and CheW.III *Architectures*) strongly suggests that the domains in a single multidomain CheW protein often belong to different sequence *Clusters* or even *Groups*, which could suggest domain specialization. On the other hand, because experimentally characterized single domain CheW proteins belonged to *Class* 1, which also includes ~93% of domains in multi-domain CheW proteins, our results suggest that there is likely little to be learned about general CheW function by exploring multi-domain CheW proteins. Presumably, multi-domain CheW proteins support formation of complexes with specific components. This conjecture is supported by the observation that CheW.III proteins form cytoplasmic double layered arrays with CheA and a special type of double domain chemoreceptor in *Vibrio cholerae*.[51]

Approximately 1% of CheW-like domains in CheW proteins were CheA-like (CheW.IC *Type*) and formed a *Class* of their own. Some members of *Class* 2 appeared to span the gap between CheA proteins and CheW/CheV proteins (Figure 4). The CheW-like domain is the most recently evolved fold in proteins that support bacterial chemotaxis.[52] The birth of CheW-like domains allowed the separation of receptor and kinase functions into different proteins, in turn facilitating the integration of multiple

stimuli sensed by arrays of different receptors into a single kinase. We are unaware of any evidence establishing whether CheW-like domains evolved first in CheW, CheA, or CheV proteins. Because CheV proteins are less common and less widely distributed than CheA or CheW proteins,[52] it seems reasonable that CheV proteins came after CheA and CheW. Because members of *Class* 2 segregated with *Class* 1, *Class* 3 and in between in our analyses (Figure 4), *Class* 2 may represent a “missing link” in the evolution of CheW-like domains from CheW to CheA or from CheA to CheW. Closer examination of sequences in *Class* 2 may provide new insights into the evolution of chemotaxis systems.

The existence of *Class* 3, which covered CheW-like domains in most CheA proteins, suggests that what is known about CheW-like domains in experimentally well-characterized CheA proteins is likely to be applicable to the majority of CheW-like domains in CheA proteins. The exceptions would be *Class* 4 and 5 CheW-like domains.

Class 4 CheW-like domains were found in CheA.VII and CheA.X *Architectures*, which have three Hpt domains and a C-terminal receiver domain and to the best of our knowledge have not been experimentally characterized. However, *Pseudomonas aeruginosa* ChpA, which has six Hpt domains but controls Type IV pili rather than flagella, supports an intramolecular phosphorelay.[53] Phosphotransfer within ChpA proceeds from ATP to the Hpt domains close to the HATPase_c domain to the receiver domain to Hpt domains further from the HATPase_c domain. Thus, a plausible hypothesis is that a distinct *Class* of CheW-like domains is required to support intramolecular phosphorelays in CheA proteins with multiple Hpt domains and a

receiver domain. Perhaps the receiver domain must be held in a particular orientation to be accessible to multiple Hpt domains.

Class 5 was the C-terminal CheW-like domain in CheA proteins with dual CheW-like domains. *A. caulinodans* CheA,[50] *Azospirillum brasilense* CheA1,[54] and *Rhodospirillum centenum* CheA1[55] have been explored experimentally and all belong to the CheA.V *Architecture* with two CheW-like domains. In each case, the C-terminal receiver domain is functionally important. However, the only experimental investigation of a *Class* 5 CheW-like domain of which we are aware was in *A. caulinodans* CheA, where removal of the C-terminal CheW-like domain diminishes chemotaxis but does not reveal the functional role of the *Class* 5 domain.[50] *Class* 5 CheW-like domains were substantially different from experimentally characterized CheW-like domains. Therefore, because investigation of *Class* 5 is especially likely to reveal new information, *Class* 5 arguably represents the highest priority for future research on CheW-like domains.

Class 6 was a *Type* (CheW.IA) encompassing ~20% of CheW proteins that appeared different from most CheW proteins (*Class* 1) in some analyses and similar in others. Method-dependent classification suggests that *Class* 6 is unlikely to represent a high priority for investigation.

4. 3 | Generally applicable methods to analyze protein domains as a function of architectural context

We have described a workflow (Figure 1) to analyze variation between sequences of a single protein domain as a function of architectural context within proteins. To ensure a robust workflow, we tested various approaches to optimize our method, including using multiple different amino acid substitution models and

multidimensional scaling techniques to estimate and visualize relevant domain sequence subpopulations. Furthermore, the similarity of our partition results using these different models was compared using ARI and NMI analyses. We first applied a basic *k*-medoid clustering step to quantify the relative homogeneity of a given domain sequence set. This allowed us to identify constituent strata for representation in a more complete sequence analysis utilizing a random stratified sub-sampling scheme. We demonstrated our approach using CheW-like domains as a general case study. The results were largely independent of methods chosen, suggesting robust conclusions. Grouping sequence *Types* of CheW-like domains into five or six major *Classes* led to a variety of useful observations, insights, hypotheses, and suggested areas of future experimental investigation.

It might appear challenging to apply our approach to domains found in many distinct protein architectures. However, protein architecture distributions tend to have a long tail, with only a few organizations representing the majority of the cases. Here, 16 out of 111 architectures in Pfam that contained CheW-like domains incorporated ~94% of CheW-like domains. Of course, individual “rare” architectures of interest could always be manually included with minimal difficulty. Another concern could involve the number of sequences to analyze. Our case study involved analysis of ~8500 sequences. Significantly more or less sequences may be available for other domains. However, our approach should still be practical. The number of starting sequences to be analyzed can be managed by choice of an appropriate input data set. We chose (and recommend) the use of Representative Proteome sets as a starting point, where the sequence space size and granularity can be modified by changing the co-membership threshold of the

relevant RP setcutoff (we used 35% as our threshold, but RP sets generated using 15%, 55%, and 75% are also available).[14] The subsampling used in our analysis (see Figure 4) is a standard strategy to deal with impractically large datasets.

Finally, we developed and have made available (see Methods) SimpLogo, a new means of visualizing amino acid sequence conservation patterns across many subtypes of the same domain and for sequences of arbitrary length. Comparison of the SimpLogos in Figure 5 and traditional sequence logos in Figure S26 quickly reveal the utility of SimpLogos. Differences in gap patterns are immediately obvious. Identifying differences in amino acid composition at individual positions is more time consuming, but feasible. Traditional sequence logos are useful to communicate information about reasonably conserved positions but of little utility for most other positions, whereas SimpLogo provides useable information about all aligned positions within a given sequence.

ACKNOWLEDGEMENTS

We thank Emily N. Kennedy and Sarah A. Barr for their insightful input and helpful discussions, including Sarah's idea that receiver domains of CheV proteins might serve as reservoirs rather than sinks for phosphoryl groups.

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. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2020;49(D1):D412-D419.
2. Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. Evolution of two-component signal transduction. *Mol Biol Evol.* 2000;17(12):1956-1970.
3. Pao GM, Saier MH, Jr. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. *J Mol Evol.* 1995;40(2):136-154.
4. Gao R, Bouillet S, Stock AM. Structural basis of response regulator function. *Annu Rev Microbiol.* 2019;73:175-197.
5. Immormino RM, Silversmith RE, Bourret RB. A variable active site residue Iinfluences the kinetics of response regulator phosphorylation and dephosphorylation. *Biochemistry.* 2016;55(39):5595-5609.
6. Page SC, Immormino RM, Miller TH, Bourret RB. Experimental analysis of functional bariation within protein families: Receiver domain autodephosphorylation kinetics. *J Bacteriol.* 2016;198(18):2483-2493.
7. Straughn PB, Vass LR, Yuan C, Kennedy EN, Foster CA, Bourret RB. Modulation of response regulator CheY reaction kinetics by two variable residues that affect conformation. *J Bacteriol.* 2020;202(15).
8. Thomas SA, Immormino RM, Bourret RB, Silversmith RE. Nonconserved active site residues modulate CheY autophosphorylation kinetics and phosphodonor preference. *Biochemistry.* 2013;52(13):2262-2273.

9. Bi S, Sourjik V. Stimulus sensing and signal processing in bacterial chemotaxis. *Curr Opin Microbiol.* 2018;45:22-29.
10. Abedrabbo S, Castellon J, Collins KD, Johnson KS, Ottemann KM. Cooperation of two distinct coupling proteins creates chemosensory network connections. *Proc Natl Acad Sci U S A.* 2017;114(11):2970-2975.
11. Huang Z, Pan X, Xu N, Guo M. Bacterial chemotaxis coupling protein: Structure, function and diversity. *Microbiol Res.* 2019;219:40-48.
12. Alexandre G, Zhulin IB. Different evolutionary constraints on chemotaxis proteins CheW and CheY revealed by heterologous expression studies and protein sequence analysis. *J Bacteriol.* 2003;185(2):544-552.
13. Gumerov VM, Andrianova EP, Zhulin IB. Diversity of bacterial chemosensory systems. *Curr Opin Microbiol.* 2021;61:42-50.
14. 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.
15. Consortium TU. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2020;49(D1):D480-D489.
16. Eddy SR. Accelerated Profile HMM Searches. *PLOS Comp Biol.* 2011;7(10):e1002195.
17. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972-1973.

18. Pelé J, Bécu J-M, Abdi H, Chabbert M. Bios2mds: An R package for comparing orthologous protein families by metric multidimensional scaling. *BMC Bioinform.* 2012;13(1):133.
19. Dayhoff M, Schwartz R, Orcutt B. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure.* 1978;5:345-352.
20. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. *Science.* 1992;256(5062):1443-1445.
21. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* 1992;89(22):10915-10919.
22. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 1992;8(3):275-282.
23. *cluster: Cluster analysis basics and extensions* [computer program]. Version 2.1.22012.
24. *fpc: Flexible procedures for clustering* [computer program]. Version 2.2-92015.
25. Hubert L, Arabie P. Comparing partitions. *J Classif.* 1985;2(1):193-218.
26. Steinley D. Properties of the Hubert-Arable Adjusted Rand Index. *Psychol Methods.* 2004;9(3):386-396.
27. Cover TM, Thomas JA. *Elements of information theory.* Wiley-Interscience; 2006.
28. Strehl A, Ghosh J. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *J Mach Learn Res.* 2003;3(null):583–617.
29. *vegan: Community Ecology Package. R package version 2.5-6. 2019* [computer program]. Version 2.5-72020.

30. Wickham H. *ggplot2: Elegant graphics for data analysis.* 3rd ed: Springer-Verlag New York; 2016.
31. *Biostrings: Efficient manipulation of biological strings* [computer program]. Version 2.60.12015.
32. Ou J, Wolfe SA, Brodsky MH, Zhu LJ. motifStack for the analysis of transcription factor binding site evolution. *Nat Methods.* 2018;15(1):8-9.
33. Dey KK, Xie D, Stephens M. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinform.* 2018;19(1):473.
34. Li X, Fleetwood AD, Bayas C, et al. The 3.2 Å resolution structure of a receptor:CheA:CheW signaling complex defines overlapping binding sites and key residue interactions within bacterial chemosensory arrays. *Biochemistry.* 2013;52(22):3852-3865.
35. Bilwes AM, Alex LA, Crane BR, Simon MI. Structure of CheA, a signal-transducing histidine kinase. *Cell.* 1999;96(1):131-141.
36. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* 2015;10(6):845-858.
37. *PyMOL* [computer program]. Version 2.3.3. http://www.pymol.org/pymol2020.
38. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53-65.
39. Tzeng J, Lu HH-S, Li W-H. Multidimensional scaling for large genomic data sets. *BMC Bioinform.* 2008;9(1):179.
40. Schneider TD, Stephens RM. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 1990;18(20):6097-6100.

41. Boukhvalova M, VanBruggen R, Stewart RC. CheA kinase and chemoreceptor interaction surfaces on CheW. *J Biol Chem.* 2002;277(26):23596-23603.
42. 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.
43. Griswold IJ, Zhou H, Matison M, et al. The solution structure and interactions of CheW from *Thermotoga maritima*. *Nat Struct.* 2002;9(2):121-125.
44. Piñas GE, Frank V, Vaknin A, Parkinson JS. The source of high signal cooperativity in bacterial chemosensory arrays. *Proc Natl Acad Sci U S A.* 2016;113(12):3335-3340.
45. Liu JD, Parkinson JS. Genetic evidence for interaction between the CheW and Tsr proteins during chemoreceptor signaling by *Escherichia coli*. *J Bacteriol.* 1991;173(16):4941-4951.
46. Porter SL, Wadhams GH, Armitage JP. *Rhodobacter sphaeroides*: complexity in chemotactic signalling. *Trends Microbiol.* 2008;16(6):251-260.
47. 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.
48. 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.

49. Karatan E, Saulmon MM, Bunn MW, Ordal GW. Phosphorylation of the response regulator CheV is required for adaptation to attractants during *Bacillus subtilis* chemotaxis. *J Biol Chem.* 2001;276(47):43618-43626.
50. Kennedy EN, Barr SA, Liu X, et al. *Azorhizobium caulinodans* chemotaxis is controlled by an unusual phosphorelay network. *J Bacteriol.* 2021:JB0052721.
51. Briegel A, Ortega DR, Mann P, Kjaer A, Ringgaard S, Jensen GJ. Chemotaxis cluster 1 proteins form cytoplasmic arrays in *Vibrio cholerae* and are stabilized by a double signaling domain receptor DosM. *Proc Natl Acad Sci U S A.* 2016;113(37):10412-10417.
52. Wuichet K, Zhulin IB. Origins and diversification of a complex signal transduction system in prokaryotes. *Sci Signal.* 2010;3(128):ra50-ra50.
53. Silversmith RE, Wang B, Fulcher NB, Wolfgang MC, Bourret RB. Phosphoryl group flow within the *Pseudomonas aeruginosa* Pil-Chp chemosensory system: Differential function of the eight phosphotransferase and three receiver domains. *J Biol Chem.* 2016;291(34):17677-17691.
54. Bible AN, Stephens BB, Ortega DR, Xie Z, Alexandre G. Function of a chemotaxis-like signal transduction pathway in modulating motility, cell clumping, and cell length in the alphaproteobacterium *Azospirillum brasilense*. *J Bacteriol.* 2008;190(19):6365-6375.
55. Jiang ZY, Gest H, Bauer CE. Chemosensory and photosensory perception in purple photosynthetic bacteria utilize common signal transduction components. *J Bacteriol.* 1997;179(18):5720-5727.

Table 1. Distribution of CheW-like domain *Context*s across *k*-medoid *Clusters*[a]

Architectural Context		Groups of Clusters																					#
		1							2	3						4							
		C	E	H	I	L	M	R	P	N	O	Q	S	T	U	A	B	D	F	G	J	K	Seqs
CheA-lineage	CheA.I	444	127	64	257	170																	1062
	CheA.II	140	312	3	2	8	167																632
	CheA.III	75	104	18	18	26	29	8								1							279
	CheA.IV	74	79	48	6	22		4															233
	CheA.V.1		4	88																			92
	CheA.V.2	2	3						89														94
	CheA.VI.1	10	3	73	2																		88
	CheA.VI.2	2	3		1				78							1	2	1					88
	CheA.VII	3	8		1			68															80
	CheA.VIII	65																					65
	CheA.IX	14	8	26	5	3		1															57
	CheA.X	5			1			50															56
	CheA.XI	29	7	2	10	3	1																52
	CheA.XII	23	22	1	1			1															48
CheV-lineage	CheV.I	4			1			1								50	11	1	4	455	290		817
CheW-lineage	CheW.I	20	4	5	11	3		2		189	157	135	117	115	92	1093	979	605	485	5	7	234	4258
	CheW.II.1	3		2	1						3	1				8	11	20	5				54
	CheW.II.2	2							1	2	1				1	12		32	2			1	54
	CheW.III.1	1								2	3				1	62	21	35	13			2	140
	CheW.III.2	1			1						3			1	1	107	4	13	8				139
	CheW.III.3	1									6					91	17	12	13				140
# Seqs		918	684	330	318	235	197	135	168	193	173	136	117	116	95	1425	1045	719	530	460	297	237	8528

[a]*Clusters* and *Groups* taken from analysis shown in Figure 3. Table entries show the number of sequences of each architectural *Context* found within each *Cluster*. *Clusters* are named in order of decreasing size.

FIGURE LEGENDS

FIGURE 1. Hierarchy of classification terms and sorting workflow. The initial steps were derived purely from protein *Architectures*, which may contain CheW-like domains in more than one *Context*. The number of *Contexts* defined (dashed arrow) the number of *Clusters* into which sequences of CheW-like domains were initially sorted. However, sequences in a given *Context* sorted into different *Clusters* and vice versa. The *Clusters* sorted into larger *Groups* of sequences. Comparison of architectural *Contexts* with sequence *Groups* revealed that all but one *Context* included *Clusters* within a single *Group*. One *Context* contained *Clusters* that spanned three *Groups* and so was split into three *Types*, whereas all other *Contexts* became a single *Type*. Minor *Clusters* of sequences associated with each *Context* were discarded from further analysis during the transition to *Types*. The *Types* sorted into larger *Classes* of sequences associated with one or more architectural *Contexts*. Figure 5 illustrates the final *Types* and *Classes* but was not part of the classification process. See text and indicated Figures or Table for details.

FIGURE 2. Schematic of the most abundant CheW-containing protein *Architectures* seen in nature. Sixteen distinct architectural variants containing the CheW-like domain were identified within the Pfam database (≥40 sequences each, ordered by decreasing abundance within each shaded group). Sequences for their corresponding CheW-like domains were extracted and initially classified according to the schematic. For parent proteins containing multiple CheW-like domains, CheW-like domains were initially classified according to the position (*Context*) of the domain, resulting in one sequence entry per CheW-like domain. Representative diagrams are

shown for each *Architecture* (from left to right, N- to C-terminal). Pfam domain IDs included are as follows: CheW-like (PF01584); H-kinase dimerization (PF02895); HPt (PF01627); Response regulatory (PF00072); CheY-binding (PF09078); HATPase_c (PF02518); P2 (PF07194). Individual schematics are not necessarily drawn to scale.

FIGURE 3. **Cluster analysis and multidimensional scaling of representative sequences of CheW-like domains reveals *Groups*.** Multidimensional scaling plots showing representative medoids from each of the 21 *Clusters*, generated using dissimilarity matrix calculated with the GONNET substitution model. Each clustering attempt reconstituted roughly the same data structure, consisting of three main *Groups* (numbered 1, 3 and 4) of *Clusters* along with a weaker, more intermediate *Group* (numbered 2) intersecting *Groups* 1 and 3. While the colors and specific *Cluster* designations may shift between analyses (overall *Cluster* sizes were slightly altered depending on the substitution matrix used, affecting *Cluster* names), the actual partitioning of CheW sequences was highly consistent, regardless of the substitution model used (see Figure S1).

FIGURE 4. Metric multidimensional scaling plots showing 25 randomly subsampled sequences from each of the 23 CheW-like domain *Types*. The top three dimensions (Dim1, Dim2, Dim3) were included for visualization purposes. Analysis was performed on dissimilarities calculated using the GONNET substitution model. The location of *Types* belonging to each of six *Classes* are indicated. Additional sampling iterations are shown in Figures S2-S4. Sampling iterations with other substitution models are shown in Figures S5-S13.

FIGURE 5. **SimpLogo representation of major CheW-like domain *Types*.** Rows correspond to domain *Types*. Sequences were aligned to the CheW-like domain HMM (Pfam ID PF01584) and separated by *Type*. Residues with physicochemical similarities were merged and assigned colors. Idealized colors (legend) were weighted by residue frequency and mixed to produce a colored box at each position. Heights were scaled by gap percentage. To differentiate between secondary colors and a near-equal mixing of primary colors, colored strips were included on the top and bottom of each box, indicating first and second most frequent residue types respectively. *Types* were organized by *Class* in the figure. Position-wise information content (in bits) was calculated for each alignment (top). Numbering is for *E. coli* CheW, framing our observations with reference to the well-studied standalone protein.

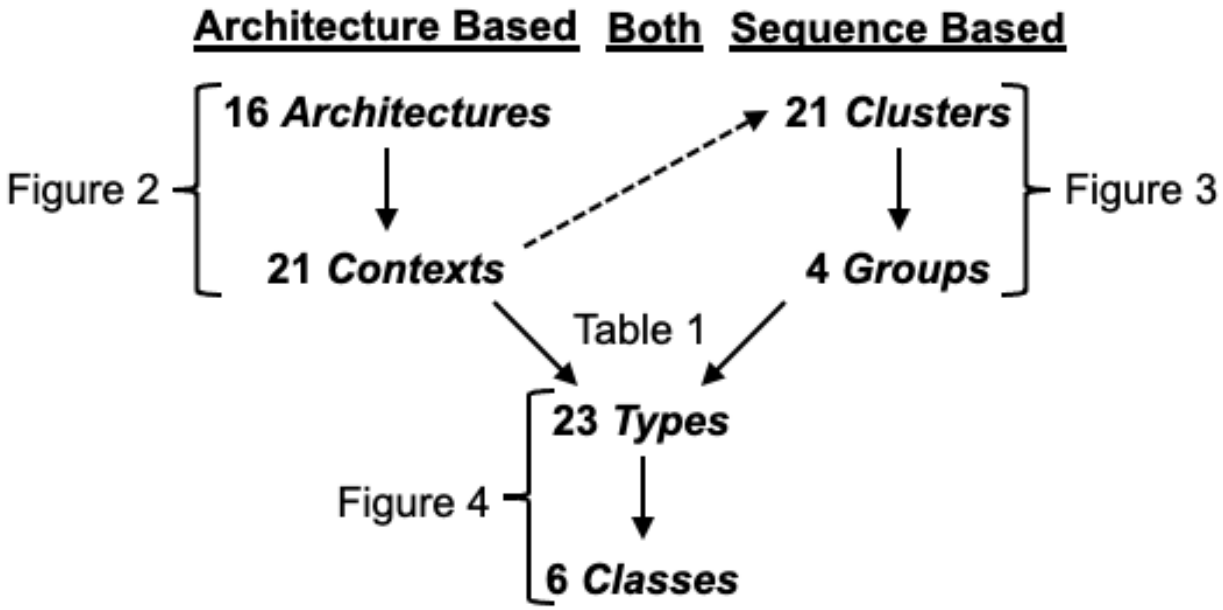


FIGURE 1. Hierarchy of classification terms and sorting workflow. The initial steps were derived purely from protein *Architectures*, which may contain CheW-like domains in more than one *Context*. The number of *Contexts* defined (dashed arrow) the number of *Clusters* into which sequences of CheW-like domains were initially sorted. However, sequences in a given *Context* sorted into different *Clusters* and vice versa. The *Clusters* sorted into larger *Groups* of sequences. Comparison of architectural *Contexts* with sequence *Groups* revealed that all but one *Context* included *Clusters* within a single *Group*. One *Context* contained *Clusters* that spanned three *Groups* and so was split into three *Types*, whereas all other *Contexts* became a single *Type*. Minor *Clusters* of sequences associated with each *Context* were discarded from further analysis during the transition to *Types*. The *Types* sorted into larger *Classes* of sequences associated with one or more architectural *Contexts*. Figure 5 illustrates the final *Types* and *Classes* but was not part of the classification process. See text and indicated Figures or Table for details.

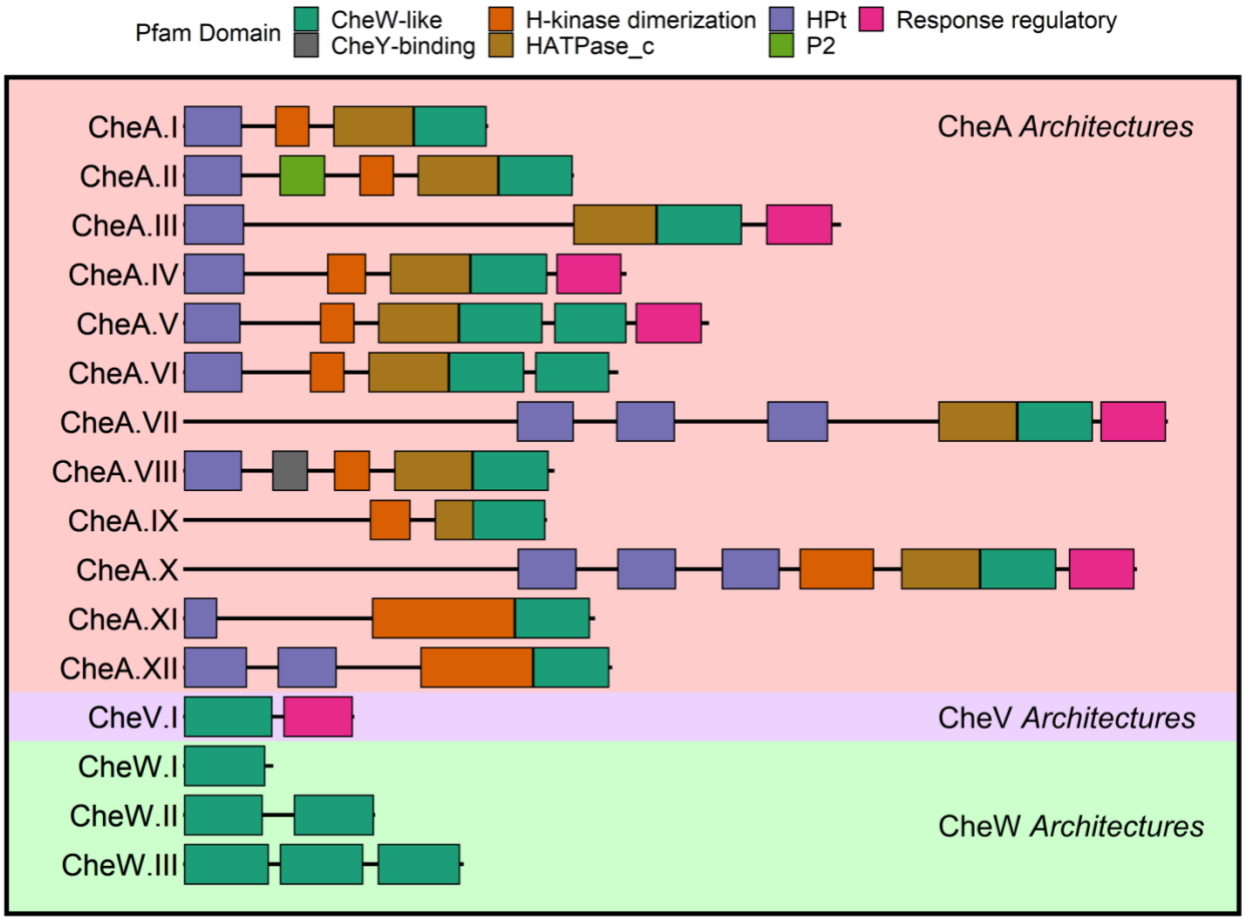


FIGURE 2. Schematic of the most abundant CheW-containing protein *Architectures* seen in nature. Sixteen distinct architectural variants containing the CheW-like domain were identified within the Pfam database (≥40 sequences each, ordered by decreasing abundance within each shaded group). Sequences for their corresponding CheW-like domains were extracted and initially classified according to the schematic. For parent proteins containing multiple CheW-like domains, CheW-like domains were initially classified according to the position (*Context*) of the domain, resulting in one sequence entry per CheW-like domain. Representative diagrams are shown for each *Architecture* (from left to right, N- to C-terminal). Pfam domain IDs included are as follows: CheW-like (PF01584); H-kinase dimerization (PF02895); HPt (PF01627); Response regulatory (PF00072); CheY-binding (PF09078); HATPase_c (PF02518); P2 (PF07194). Individual schematics are not necessarily drawn to scale.

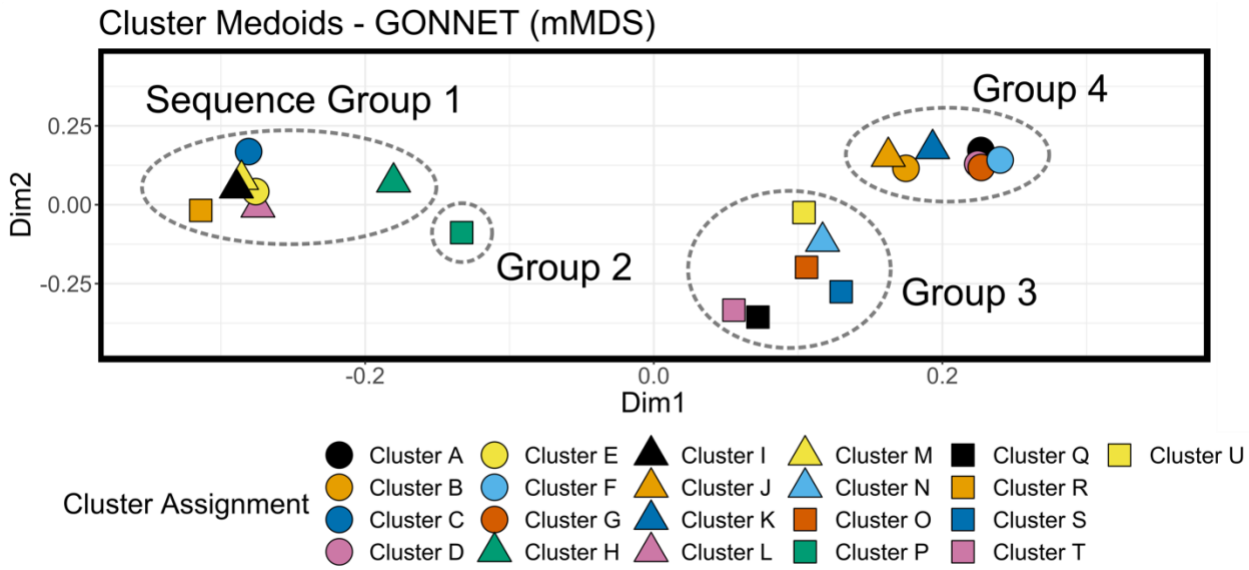


FIGURE 3. **Cluster analysis and multidimensional scaling of representative sequences of CheW-like domains reveals *Groups*.** Multidimensional scaling plots showing representative medoids from each of the 21 *Clusters*, generated using dissimilarity matrix calculated with the GONNET substitution model. Each clustering attempt reconstituted roughly the same data structure, consisting of three main *Groups* (numbered 1, 3 and 4) of *Clusters* along with a weaker, more intermediate *Group* (numbered 2) intersecting *Groups* 1 and 3. While the colors and specific *Cluster* designations may shift between analyses (overall *Cluster* sizes were slightly altered depending on the substitution matrix used, affecting *Cluster* names), the actual partitioning of CheW sequences was highly consistent, regardless of the substitution model used (see Figure S1).

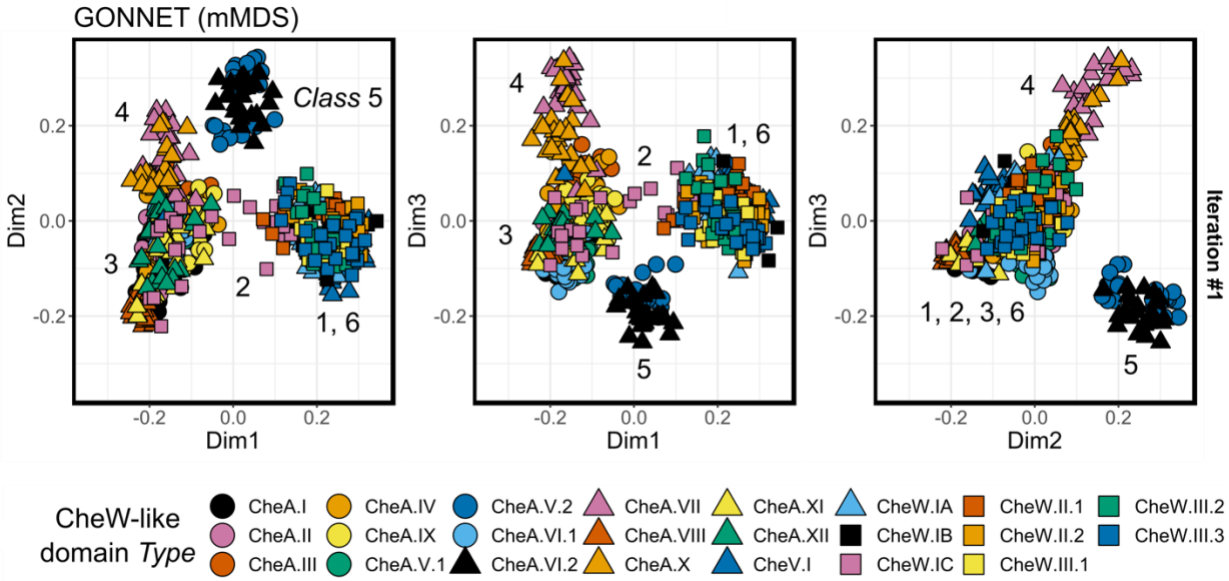
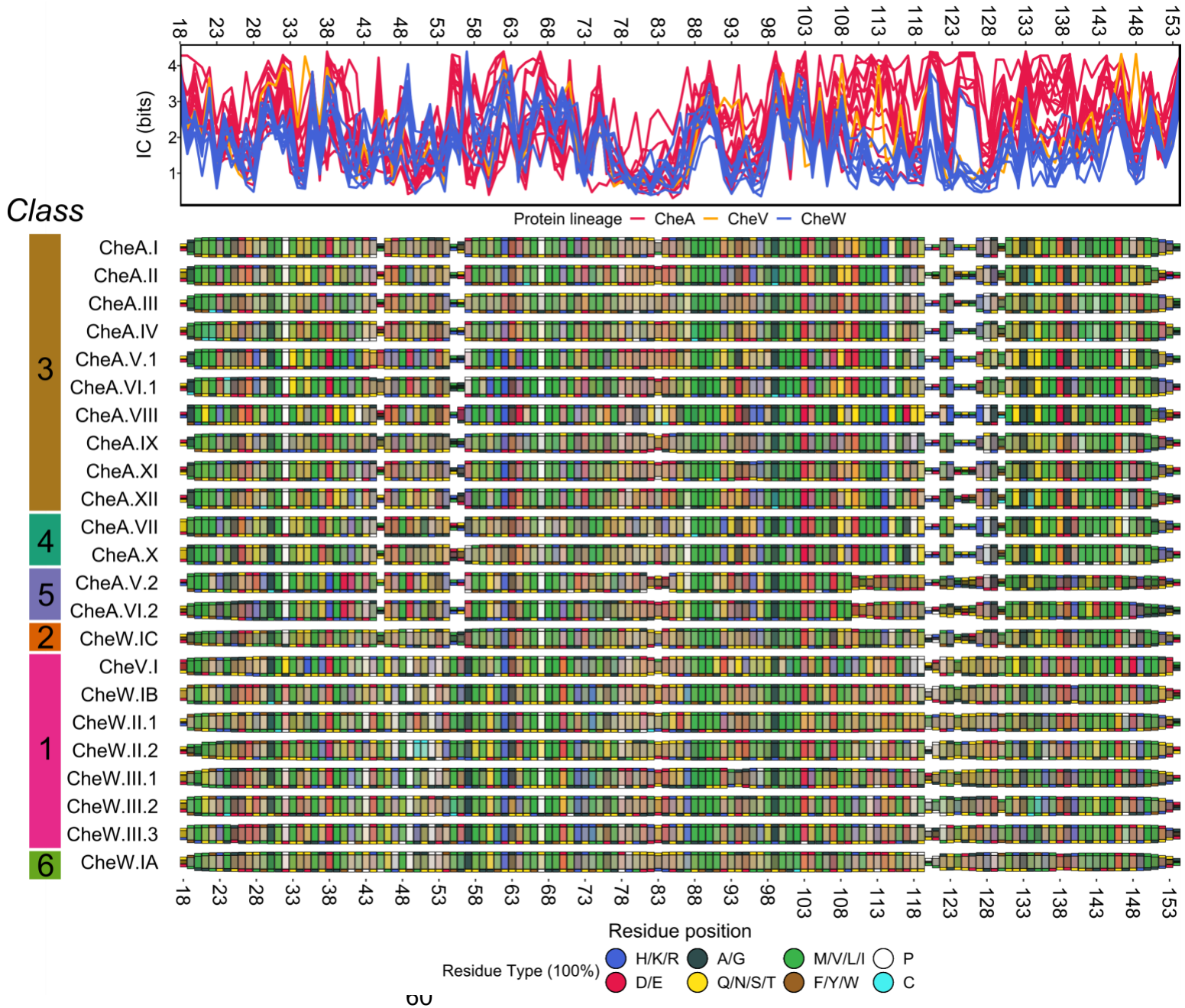


FIGURE 4. Metric multidimensional scaling plots showing 25 randomly subsampled sequences from each of the 23 CheW-like domain *Types*. The top three dimensions (Dim1, Dim2, Dim3) were included for visualization purposes. Analysis was performed on dissimilarities calculated using the GONNET substitution model. The location of *Types* belonging to each of six *Classes* are indicated. Additional sampling iterations are shown in Figures S2-S4. Sampling iterations with other substitution models are shown in Figures S5-S13.


IC (bits)
Protein lineage CheA CheV CheW
Class
CheA.I
CheA.II
CheA.III
CheA.IV
CheA.V.1
CheA.VI.1
CheA.VIII
CheA.IX
CheA.XI
CheA.XII
CheA.VII
CheA.X
CheA.V.2
CheA.VI.2
CheW.IC
CheV.I
CheW.IB
CheW.II.1
CheW.II.2
CheW.III.1
CheW.III.2
CheW.III.3
CheW.IA
Residue position
Residue Type (100%)
H/K/R
D/E
A/G
Q/N/S/T
M/V/L/I
F/Y/W
P
C

1175

FIGURE 5. **SimpLogo representation of major CheW-like domain *Types*.** Rows correspond to domain *Types*. Sequences were aligned to the CheW-like domain HMM (Pfam ID PF01584) and separated by *Type*. Residues with physicochemical similarities were merged and assigned colors. Idealized colors (legend) were weighted by residue frequency and mixed to produce a colored box at each position. Heights were scaled by gap percentage. To differentiate between secondary colors and a near-equal mixing of primary colors, colored strips were included on the top and bottom of each box, indicating first and second most frequent residue types respectively. *Types* were organized by *Class* in the figure. Position-wise information content (in bits) was calculated for each alignment (top). Numbering is for *E. coli* CheW, framing our observations with reference to the well-studied standalone protein.