Biochemistry

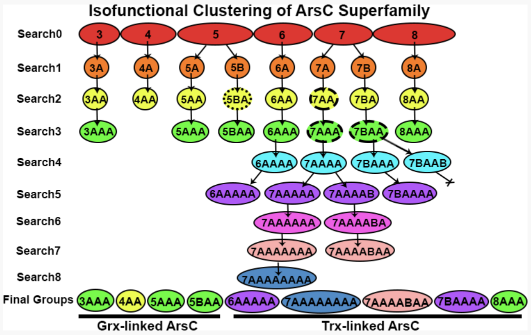
pubs.acs.org/biochemistry Article

Isofunctional Clustering and Conformational Analysis of the Arsenate Reductase Superfamily Reveals Nine Distinct Clusters

Mikaela R. Rosen, Janelle B. Leuthaeuser, Carol A. Parish,* and Jacquelyn S. Fetrow*

Cite This: *Biochemistry* 2020, 59, 4262−4284 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Arsenate reductase (ArsC) is a superfamily of enzymes that reduce arsenate. Due to active site similarities, some ArsC can function as low-molecular weight protein tyrosine phosphatases (LMW-PTPs). Broad superfamily classifications align with redox partners (Trx- or Grx-linked). To understand this superfamily's mechanistic diversity, the ArsC superfamily is classified on the basis of active site features utilizing the tools TuLIP (two-level iterative clustering process) and autoMISST (automated multilevel iterative sequence searching technique). This approach identified nine functionally relevant (perhaps isofunctional) protein groups. Five groups exhibit distinct ArsC mechanisms. Three are Grx-linked: group 4AA (classical ArsC), group 3AAA (YffB-like), and group 5BAA. Two are Trx-linked: groups 6AAAAA and 7AAAAAAAA. One is an Spx-like transcriptional regulatory group, group 5AAA. Three are potential LMW-PTP groups: groups 7BAAAA, and 7AAAABAA, which have not been previously identified, and the well-studied LMW-PTP family group 8AAA. Molecular dynamics simulations were utilized to explore functional site details. In several families, we confirm and add detail to literature-based mechanistic information. Mechanistic roles are hypothesized for conserved active site residues in several families. In three families, simulations of the unliganded structure sample specific conformational ensembles, which are proposed to represent either a more ligand-binding-competent conformation or a pathway toward a more binding-competent state; these active sites may be designed to traverse high-energy barriers to the lower-energy conformations necessary to more readily bind ligands. This more detailed biochemical understanding of ArsC and ArsC-like PTP mechanisms opens possibilities for further understanding of arsenate bioremediation and the LMW-PTP mechanism.

The arsenate reductase (ArsC) superfamily is a large superfamily of proteins involved in the metabolism of arsenic. Arsenic concentrations have been increasing, especially due to the increased level of pollutants resulting from mining and agricultural activities.[1] Arsenic, a highly toxic and carcinogenic metalloid, is dangerous to the body because, in its various redox forms, it can mimic and potentially replace phosphates in molecular reactions. Arsenate can replace phosphate in several stages of cellular respiration, and arsenite can inhibit pyruvate dehydrogenase.[1] ArsC proteins are vital to the function of arsenic redox microorganisms that decrease the concentrations of arsenate. Therefore, a more complete understanding of enzymatic mechanisms within the ArsC superfamily is important to bioremediation strategies that prevent arsenic from reaching alarming levels in the environment.

The ArsC protein superfamily is a member of the thioredoxin (Trx)-fold family, a very large and ubiquitous family known for its characteristic four-stranded mixed β sheet sandwiched by three α helices, as well as a reactive cysteine at the active site and many variations on a CXXC motif.[2−4] According to Atkinson and Babbitt, the ArsC superfamily is unlike other superfamilies in the Trx-fold family.[5] The network-based classification system on full protein sequences used by Atkinson and Babbitt suggests that, potentially, the arsenate reductases make up one superfamily that might be broken into three subfamilies. While this full sequence network-based approach provides good foundational information, it is not sensitive enough to easily subdivide superfamilies into their functionally relevant families on the basis of active site details.[6]

Although researchers have focused on specific subfamilies within the arsenate reductase superfamily, the breadth of the ArsC superfamily is largely unstudied. Research that explores arsenate subfamilies includes (but is not limited to) glutaredoxin (Grx)-coupled, thioredoxin (Trx)-coupled, and Acr2 (also known as eukaryotic ArsC).[7] The first two classifications are based on cellular redox partners. The second is based on organismal (e.g., eukaryotic) distribution. Less robust ArsC mechanisms being researched include mycothiol

Received: August 5, 2020
Revised: October 1, 2020
Published: November 2, 2020

and trypanothione-linked, which suggests a complex pathway in which an organism will utilize two different types of ArsC to function most efficiently.[8] A hybrid Grx/Trx-linked mechanism has also been described in which the ArsC structurally resembles the Trx-coupled ArsC, but relies on Grx for its reactivation.[9] In addition, a known family of ArsC-related proteins is a distinct class of protein tyrosine phosphatases (PTPs) called the low-molecular weight phosphatases (LMW-PTP), some of which can bind both phosphate and arsenate.[10,11] A comparison of the proposed enzymatic mechanisms of Grx-linked ArsCs, Trx-linked ArsCs, and LMW-PTPs shows the commonalities and what is currently known about the differences among them (Figure S1). Both ArsC and LMW-PTP proteins form cysteine intermediates (arsenocysteine and phosphocysteine); however, the ArsC proteins require the action of disulfide cascades and interaction with redox partners, while the LMW-PTP simply undergoes hydrolysis. These classifications are based on understanding of cell biology and the identity of redox partners. As new proteins are discovered through the genome sequencing projects, understanding the relationship between the fold and active site mechanism is essential to understanding the breadth of mechanisms that can perform a given function.

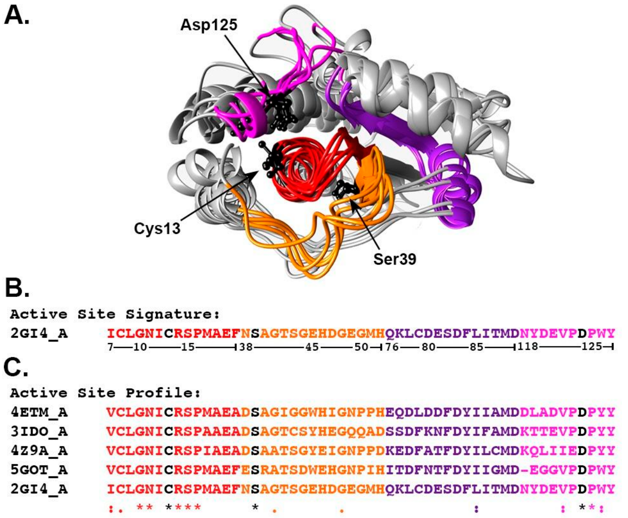
In this work, our goal was to broadly evaluate the ArsC superfamily and identify distinct functionally relevant families. In the ideal case, functionally relevant clusters would be isofunctional, meaning all proteins in the cluster would share the same function at the level of the biochemical mechanism. Besides our own long-standing research, this approach has also been explored by others.[12−14] The "active (or functional) site signature" is a key feature of these approaches. Such a signature consists of the mechanistic or functional determinants, just the residues that are specifically involved in the enzyme function being classified. These signatures have been called "active site signatures"[15] and "signature positions",[16] among other terms.[12]

Specifically, in our active site profiling approach, amino acid residues within 10 Å of the known functional site are identified, aligned, and compared (Figure 1).[15,17] These sequence fragments are aligned to create an active site signature for a given protein (Figure 1B, for LMW-PTP protein 2GI4). Signatures are aligned and compared to the signatures of other proteins within a family to create an active site profile for a given family (Figure 1C). This concept has been applied to multiple protein families[17−19] and was reviewed in a recent edited volume.[6]

The concept of active site profiling has been implemented in the program DASP (Deacon Active Site Profiling). DASP utilizes active site profiling to search the sequences of proteins of known structure or of GenBank.[20,21] DASP utilizes an active site profile to identify additional sequences with similar active site features. More recently, DASP was upgraded to DASP3, which was utilized in the autoMISST searches described here.[22]

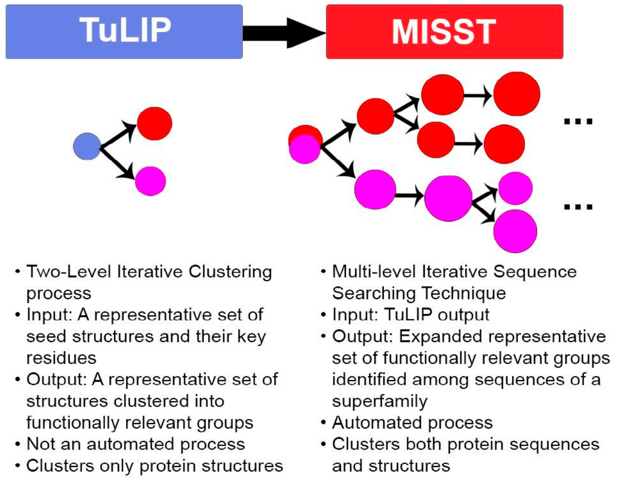
DASP3 and the concept of active site profiling have been implemented into a two-step process to identify potential functional families within a superfamily (Figure 2). The first step, TuLIP (two-level iterative clustering process), identifies functional families within proteins of known structure.[18]

These families are taken as seeds into MISST (multilevel iterative sequence searching technique), which is an iterative process that involves searching GenBank sequences for those with similar functional features. MISST differs from other methods in that it is both agglomerative and divisive. It is agglomerative because it starts with just a few groups of proteins of known structure. The process then identifies new



```
Active Site Signature:
2GI4_A     ICLGNICRSPMAEFNSAGTSGEHDGEGMHQKLCDESDFLITMDNYDEVPDPWY
           7—10———15————138—————45————50—176——80————85————1118——————125—1
Active Site Profile:
4ETM_A     VCLGNICRSPMAEADSAGIGGWHIGNPPHEQDLDDFDYIIAMDDLADVPDPYY
3IDO_A     VCLGNICRSPAAEADSAGTCSYHEGQQADSSDFKNFDYIFAMDKTTEVPDPYY
4Z9A_A     VCLGNICRSPIAEADSAATSGYEIGNPPDKEDFATFDYILCMDKQLIIEDPYY
5GOT_A     VCLGNICRSPMAEFESRATSDWEHGNPIHITDFNTFDYIIGMD-EGGVPDPWY
2GI4_A     ICLGNICRSPMAEFNSAGTSGEHDGEGMHQKLCDESDFLITMDNYDEVPDPWY
           :. ** ****     * .       .                :        : **:
```

Figure 1. Concept of active site signatures and profiles, for extraction of structural information into sequence space. (A) Structures of five ArsC-like proteins that were the founding members of group 2BABAAAA, which led to the development of a putatively isofunctional group, group 8AAA, identified in this research. Key residues are shown as black ball-and-stick models; residues withan atom within 10 Å of the center of geometry of these key residues are shown as colored ribbons. (B) Active site signature, which is created using residues within 10 Å of the key residues, for 2GI4. Each color represents a sequence fragment; residues correspond to the ribbon color in the structure. The residue numbering is shown for 2GI4, the representative of group 8AAA. (C) Several active site signatures for other members of the group are aligned to create the profile. The similarity of signatures within a profile can be quantified with an active site profile score.[15] The symbols below the alignment signify residue conservation. The active site profile is highlighted using colors (red, pink, purple, and orange) from panel A.



- Two-Level Iterative Clustering process
- Input: A representative set of seed structures and their key residues
- Output: A representative set of structures clustered into functionally relevant groups
- Not an automated process
- Clusters only protein structures

- Multi-level Iterative Sequence Searching Technique
- Input: TuLIP output
- Output: Expanded representative set of functionally relevant groups identified among sequences of a superfamily
- Automated process
- Clusters both protein sequences and structures

Figure 2. Overview of TuLIP and MISST processes. TuLIP and MISST are iterative processes that utilize DASP and DASP3, respectively, to search the RCSB structure database (TuLIP) and the GenBank sequence database (MISST) to identify protein sequences with similar active site features. During each iteration, new sequences may be added to each group (agglomerative) and the group is evaluated for the presence of multiple functionally relevant groups (divisive).

proteins with similar functional sites through iterative searches of GenBank using DASP3.[22] Because it is agglomerative, it can identify all members within a functional family whose features are sufficiently similar to meet the DASP3 score threshold on one of the superfamily's active site profiles.

MISST is divisive because, at each search iteration, heuristics are applied to identify if or when each group contains more than one functional family. If more than one functional family is identified, the script subdivides the family (heuristics described in Supplemental Methods; no user input is required during autoMISST). To determine subdivision, MISST uses the observation that DASP3 searches self-identify isofunctional groups, but groups that include more broad function do not self-identify.[18] After division, an active site profile is created for each new and putatively isofunctional group. Each, individually, is then used to search GenBank during the next MISST iteration. Thus, a branch of the superfamily in which the active site details are too divergent from any of the superfamily's active site profiles would be missed. MISST has been implemented in a fully automated process, autoMISST,[23] which (in its unautomated form) has been successfully applied to the peroxiredoxin superfamily.[19]

How does MISST differ from other methods for identifying potentially functionally relevant clusters? It differs from the vast majority of such methods, including the sequence similarly network approach previously used to identify three clusters within the ArsC superfamily,[5] by focusing on only details around the functional site, instead of using the full length sequence. MISST thus focuses on these functional details, which may get lost within the entire protein sequence. The families identified are focused on molecular function, rather than a broader cellular or physiological function.[6] Two approaches most similar to MISST are work from the De Melo−Minardi laboratory based on ASMC[13,14] and applications from the Orengo laboratory based on FunHMMer[12,24,25] ASMC-based approaches that identify functionally relevant groups within PFAM[26] families by first building comparative models for each sequence, then using genetic programming to identify putative functional features, and then subdividing the PFAM family on the basis of those features. The Orengo laboratory approach is to use CD-Hit[27,28] to search for sequences in known functional families and then identify active site features and mutations that might correlate with function. MISST is the only approach that is both agglomerative, searching all of GenBank iteratively, and divisive, identifying when one group contains potentially two functionally relevant clusters. When such a split occurs, each cluster is used to search GenBank sequences, so MISST is also agglomerative for each cluster. Additionally, this process identifies when such clusters remerge, suggesting that the clusters are not completely distinct.

It is important to recognize that proteins can have multiple functional sites.[29,30] autoMISST focuses only on the local features of a single given active site; thus, a protein superfamily could be clustered multiple ways, depending on which functional site is being evaluated.

Our long-term goal has been to cluster proteins and protein superfamilies into isofunctional groups. In the work described herein, TuLIP and autoMISST were applied to the ArsC superfamily of proteins to identify the breadth of active site mechanisms in this superfamily and groups that are potentially isofunctional. TuLIP search and clustering of the proteins of known structure identified two groups of 10 nonredundant arsenate reductases, which were used as input into autoMISST. The first round of autoMISST, in turn, identified six putatively isofunctional families composed of 57729 arsenate reductase sequences through eight iterative searches of GenBank. Using the six groups output by the first round as input, the second round of autoMISST identified nine functionally relevant clusters composed of 110549 ArsC sequences through nine iterative searches of GenBank. In this contribution, these nine clusters are described as functionally relevant clusters or potential/putative isofunctional groups (representing one enzyme mechanism, potentially with one substrate specificity). We hypothesize that each group is isofunctional, but this must be evaluated by experiment.

Molecular dynamics (MD) evaluation of conserved active site residues in each family suggests that these nine potentially isofunctional groups comprise the ArsC superfamily and that each is distinguished by a distinct enzymatic mechanism. In addition, for several functionally relevant groups, the MD simulations suggest pathways for exploring ligand-competent conformations that are different for each group, and correlate with suggested high-energy transit conformations previously described.[31] Ultimately, the hypothesis that the ArsC superfamily clusters into these nine functionally relevant families provides the foundation for experimental work into enzymatic mechanisms that would potentially enable arsenate bioremediation strategies and support the growing field of LMW-PTP drug discovery motivated by the many diseases and disorders related to LMW-PTPs and PTPs, including cancer, neurodevelopmental disease, and diabetes.[10]

■ METHODS

DASP, a Method for Identifying Sequences with Functional Site Features in Common with Those Found in the Search Profile. Functionally relevant clusters in the ArsC superfamily were identified using methods previously described: TuLIP[18] and MISST.[19] The foundation of both methods is software called DASP (Deacon Active Site Profiler).[17,20−22] DASP is built on the concept of active site profiling, originally described by Cammer and colleagues,[15] which creates signatures on the basis of only the residue information in the vicinity of the active site. As described in Figure 1, signatures are comprised of residue fragments. Each fragment has more than three residues. In addition, at least one atom in each residue in each fragment is found within 10 Å of the center of geometry of functionally important "key residues" (Figure 1A). Fragments that meet this definition are concatenated to create a signature (Figure 1B), and then signatures are aligned to create an active site profile (Figure 1C).

DASP uses these profiles to search sequence databases (sequences in either the RCSB Protein Data Bank or GenBank) to identify proteins containing similar functional site features. Focusing on the sequence in the vicinity of the functional site allows DASP to identify potentially important functional determinants (the "functional signal") from the details in the broader sequence (the "noise").[15,17,32] The DASP search process involves creating a position-specific scoring matrix (PSSM) for each fragment in a given profile (e.g., a PSSM for the fragments colored red in Figure 1C). Each fragment-specific PSSM in a given profile is compared to a query sequence, one GenBank sequence, in a sliding window approach. For each alignment of the PSSM to a segment of the query sequence, a p value of "match" between the query segment and the PSSM is calculated. When each query segment has been aligned, one is identified or "tagged" as the "best match", that segment of the query sequence with the most significant *p* value match to the fragment-specific PSSM. The sliding alignment is repeated for each fragment-specific PSSM in the profile (e.g., this would be done for each of the red, orange, purple, and magenta fragments in Figure 1C). This process results in a set of "tagged" segments from the query sequence, each of which aligns best to one of the fragments in the profile. *p* value scores for all tagged segments are combined using QFAST[33] to calculate a score that represents the significance of the query sequence "matching" the active site profile used for the search. This is the DASP (or DASP3) score. This sliding alignment for each fragment-specific PSSM in the profile is completed for every sequence in GenBank; consequently, the end result is a DASP score representing the significance of each sequence in GenBank aligning to the active site profile.

In the current work, DASP[17,20,21] was utilized for the TuLIP searches. An improved version of DASP, DASP3,[22] was utilized in the autoMISST process. The score threshold that represents a significant match between a profile and a sequence has been thoroughly evaluated and depends on the program version. Relevant to this work, using DASP,[20,21] RCSB score and GenBank score thresholds are 1×10^{-12}.[17,18] Using the newer version, DASP3,[22] the RCSB score threshold is 1×10^{-14} and the GenBank score threshold is 1×10^{-16}.[19,22] Importantly, these score thresholds do not vary much between protein superfamilies.

TuLIP and MISST, Methods That Utilize DASP to Identify Functionally Relevant Clusters of Proteins. Putatively isofunctional groups in the arsenate reductase superfamily were identified using methods previously described: TuLIP (two-level iterative process[18]) and MISST (multilevel iterative sequence searching technique[17,19,23]). For the work described here, MISST was implemented as an automated process called autoMISST that requires an input of only proteins and their key residues, in either one or more functionally relevant groups.[23] autoMISST includes no other user-settable heuristics or parameters.

To begin the process, a small group of ArsC of known structure were selected from RSCB[34] and PFAM[26] to represent the diversity among structures of the known ArsC superfamily. These structures are listed in Figure 3A. To identify potential additional structures, a BLAST[35] sequence search was performed. From the group thus identified, trivial sequence redundancy was eliminated using a cutoff of 95% sequence identity. For this work, 10 "seed" proteins were identified, listed in Figure 3A (group 0), and used as input to TuLIP.

For each protein, three "key residues", residues known or thought to be functionally important, were identified from a literature review. The invariant active site Cys, known to be crucial to function in all Trx-fold family proteins,[2−5,36] is one of the three key residues identified in all proteins for this work. The other residues were identified by literature review, conservation, and structural alignment.

The 10 seed proteins and their key residues were input into TuLIP,[18] a previously described manual process that follows the following steps: (1) create an all-by-all network of the 10 proteins, in which each edge represents the pairwise signature alignment score between two proteins, (2) MCL (Markov) clustering[37,38] performed iteratively with a successively more stringent pairwise edge score to identify closely related clusters first, as illustrated by Leuthaeuser and colleagues,[32] and (3) at each MCL cluster level, determine if any self-contained clusters of three or more proteins are identified. If so, then calculate the functional relevance by determining if the cluster self-identifies in a DASP search[20−22] of RCSB Protein Data Bank protein sequences.[34,39] If a cluster or subcluster of three or more proteins self-identifies, it is deemed a functionally relevant group[18] and is set aside for input into MISST. If a cluster does not self-identify, then MCL clustering is repeated at the next more stringent edge score. Steps 1−3 are repeated until identified proteins are in functionally relevant groups, and/or are singlets. At the time this work was completed, TuLIP identified two functionally relevant groups of ArsC of known structure.

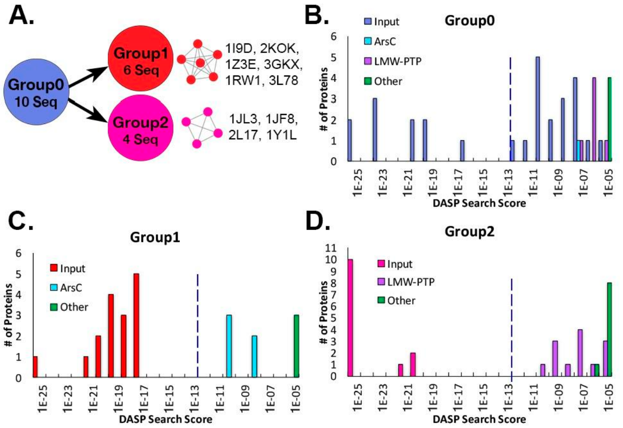
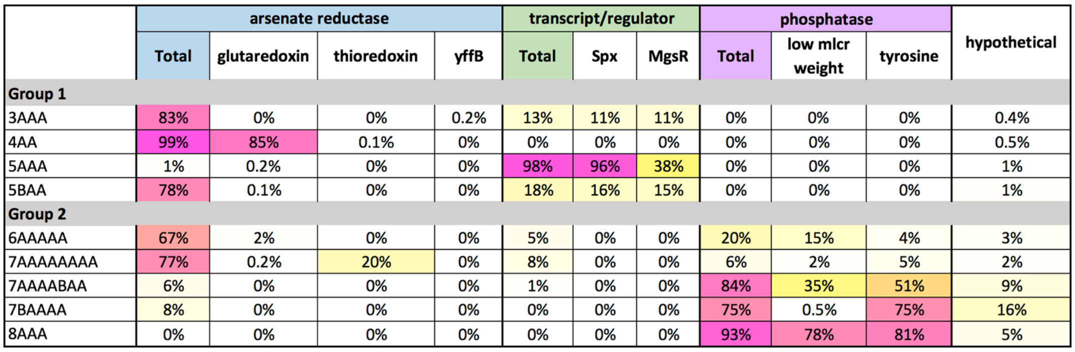


Figure 3. Iterative TuLIP applied to ArsC superfamily. (A) Two TuLIP iterations identified two functionally relevant clusters. The input (blue) included 10 seed proteins (group 0) identified through a search of the RCSB PDB, PFAM, and the literature. Two output groups (red and pink) contained six and four nonredundant proteins, respectively. (B) DASP search of RCSB PDB sequences using group 0 as input identified 10 input sequences along with 18 others trivially identical to the input sequences (blue bars). Eleven sequences not trivially identical to input sequences are colored on the basis of annotation in the PDB as ArsC (aqua bars), LMW-PTPs (purple bars), or other (green bars). The dashed vertical line indicates the 1 × 10^{-12} DASP score threshold.[18] Because group 0 input proteins did not self-identify, a second iteration was completed, producing two groups, 1 and 2. (C) Score distribution from a DASP search using the six group 1 input sequences identified the input sequences and 10 additional trivially similar sequences (red bars). Eight new proteins, scoring below the significance threshold, were identified (aqua and green bars). (D) Score distribution from a DASP search using four nonredundant group 2 proteins as input. Pink bars show these proteins and the nine proteins trivially similar to them. A total of 22 newly identified proteins, which score less significantly than 1×10^{-12}, were identified (purple and green bars). Because each group self-identified, groups 1 and 2 are functionally relevant clusters and are used as the MISST input (proteins listed in panel A).

The proteins and their key residues for each of the two TuLIP-identified potentially isofunctional groups were the input into the first round (search 0) of autoMISST (Table S1). Key residue selection for the second round of autoMISST is described subsequently.

autoMISST uses iterative DASP3 searches of GenBankNR[23] and a set of automated scripts and parameters for combining and dividing groups into functionally relevant clusters. No user-defined parametrization occurs during the autoMISST process. In the iterative autoMISST process, the DASP3 score threshold for the first search (search 0) is 1×10^{-14} and is

Table 1. GenBank Annotations for Each of the Nine Groups Identified by autoMISST[a]



	arsenate reductase				transcript/regulator			phosphatase			hypothetical
	Total	glutaredoxin	thioredoxin	yffB	Total	Spx	MgsR	Total	low mlcr weight	tyrosine	
Group 1											
3AAA	83%	0%	0%	0.2%	13%	11%	11%	0%	0%	0%	0.4%
4AA	99%	85%	0.1%	0%	0%	0%	0%	0%	0%	0%	0.5%
5AAA	1%	0.2%	0%	0%	98%	96%	38%	0%	0%	0%	1%
5BAA	78%	0.1%	0%	0%	18%	16%	15%	0%	0%	0%	1%
Group 2											
6AAAAA	67%	2%	0%	0%	5%	0%	0%	20%	15%	4%	3%
7AAAAAAAA	77%	0.2%	20%	0%	8%	0%	0%	6%	2%	5%	2%
7AAAABAA	6%	0%	0%	0%	1%	0%	0%	84%	35%	51%	9%
7BAAAA	8%	0%	0%	0%	0%	0%	0%	75%	0.5%	75%	16%
8AAA	0%	0%	0%	0%	0%	0%	0%	93%	78%	81%	5%

[a] Headers are colored to highlight general annotations as arsenate reductase (blue), transcriptional regulator (green), and phosphatase (purple). The second columns under each colored heading identify more specific subannotations that were found by searching GenBank annotation for the specific word in the heading. For example, glutaredoxin is a specific type of ArsC distinction and includes ArsC proteins annotated as Grx-linked. The total column for each general annotation corresponds to the percent annotation for that particular keyword and will not necessarily equal the sum of its more detailed subannotations.

increased to 1 × 10^{-16} for search 1. The threshold remains 1 × 10^{-16} for all subsequent searches. For the results described here, two GenBank versions were used. The first round of autoMISST was completed in March 2017 using GenBankNR version 218; the second round of autoMISST was completed in July 2019 using GenBankNR version 232. MISST and autoMISST have been validated on multiple superfamilies, including the enolase and peroxiredoxin superfamilies.[19,23]

The manuscript describing autoMISST is in process; thus, the flowchart for the automated approach and script details are described in Supplemental Methods.

Selecting Proteins and Key Residues for the Second Round of AutoMISST. The first round of autoMISST identified six functionally relevant groups. Sequences of several of these groups overlapped more than previously observed for other superfamilies, suggesting that the functionally relevant clustering was not complete (described in Supplemental Result 1). Thus, a second round of autoMISST was accomplished. We did not input all sequences from each of the six groups. Instead, only a subset of known structures from each group were input into the second round of autoMISST.

The proteins for the second round of autoMISST were selected on the basis of their significance score within their respective groups. Trivial redundancy was removed by selecting only one representative of a cluster of sequences that are >95% identical in sequence. In ArsC, one exception was a case in which all structures identified in the first round of autoMISST were similar (>95% identical); thus, all were used.

These guidelines resulted in selection of the following proteins. For group 1AAA, two structures were output from round 1 and their sequences were used as input for round 2. For group 1BAAAA, 13 structures were identified; only the three nonredundant structures were used as input. For group 1BABAAAA, six unique structures were output from round 1; all of the sequences were input into round 2. For group 2AAAAA, eight structures were output from round 1; sequences of the two nonredundant structures were input into round 2. For group 2BAAAA, seven structures were output from round 1; sequences of three nonredundant proteins were input to round 2. For group 2BABAAAA, 28 PDB entries scored more significantly than 1 × 10^{-16}. To achieve some consistency in the number of inputs for each group, a distribution of four of the most significantly scoring proteins (all better than 1 × 10^{-20}) was selected as inputs for the second round of autoMISST.

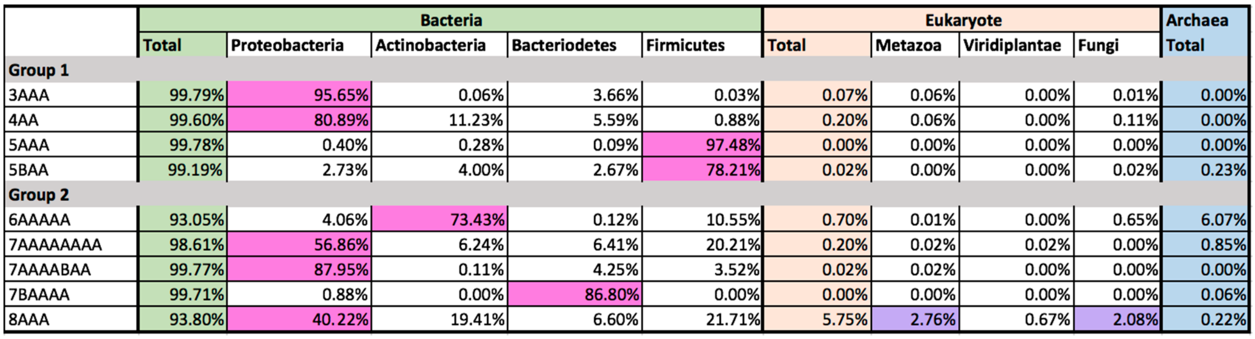
Together with the protein sequences, autoMISST requires key residues for each group. Key residues were selected on the basis of the observed first-round residue conservation and on how well the conserved residues overlaid in the structures. Key residues in the three group 1-derived groups were unchanged from the first round. Residue conservation for the group 2-derived groups suggested that new key residues should be selected. Key residues were uniformly chosen across the group 2-derived groups: the catalytic Cys, a conserved Ser in the S-A/G-G motif, and the highly conserved Asp in the DP motif. (Selected proteins and their key residues for second-round input are listed in Table S1.)

In this second round of autoMISST with six input groups, nine output groups resulted. Interestingly, groups that were not overlapping in the first round did not split further. Two groups that did overlap split in this second round, as described in the Results. Details are described in Supplemental Results 1 and 2.

Analysis of the Groups (WebLogo, Annotations). autoMISST produces sequences identified as being part of an isofunctional group (Supplemental File 1). Signature logos for each of the putatively isofunctional groups were created to determine residue conservation. For this analysis, each group included proteins that were identified with p values down to 1 × 10^{-16}, including cross-hits and redundancies. The MatchingSubSeqs, each of which represents the 10 Å around an inputted key residue, were concatenated using Microsoft Excel from the N to C terminus. Those concatenated sequence fragments, termed active site signatures, were input into WebLogo version 3.7.[40] The larger residues are more conserved within a group. In this way, we can compare residue conservation between groups and determine unique residues belonging to one group or a few.

Another analysis that was performed was based on the current documentation of the proteins in the RCSB Protein Data Bank (PDB) and GenBank. Both the current protein annotations and phylogeny were investigated. autoMISST automatically produces a file containing the annotation and phylogeny information for each protein structure or sequence identified (Supplemental File 1). This file was searched for keywords such as "arsenate" or "eukaryota" to create percentages for each annotation and phylogeny (Tables 1

Table 2. Phylogenetic Distribution of Each of the Nine Groups Identified by autoMISST



	Bacteria					Eukaryote				Archaea
	Total	Proteobacteria	Actinobacteria	Bacteriodetes	Firmicutes	Total	Metazoa	Viridiplantae	Fungi	Total
Group 1										
3AAA	99.79%	95.65%	0.06%	3.66%	0.03%	0.07%	0.06%	0.00%	0.01%	0.00%
4AA	99.60%	80.89%	11.23%	5.59%	0.88%	0.20%	0.06%	0.00%	0.11%	0.00%
5AAA	99.78%	0.40%	0.28%	0.09%	97.48%	0.00%	0.00%	0.00%	0.00%	0.00%
5BAA	99.19%	2.73%	4.00%	2.67%	78.21%	0.02%	0.00%	0.00%	0.02%	0.23%
Group 2										
6AAAAA	93.05%	4.06%	73.43%	0.12%	10.55%	0.70%	0.01%	0.00%	0.65%	6.07%
7AAAAAAAA	98.61%	56.86%	6.24%	6.41%	20.21%	0.20%	0.02%	0.02%	0.00%	0.85%
7AAAABAA	99.77%	87.95%	0.11%	4.25%	3.52%	0.02%	0.02%	0.00%	0.00%	0.00%
7BAAAA	99.71%	0.88%	0.00%	86.80%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%
8AAA	93.80%	40.22%	19.41%	6.60%	21.71%	5.75%	2.76%	0.67%	2.08%	0.22%

and 2). This allows us to see if there is an agreed upon known function (high percentage of a certain annotation) that can help identify the groupings output by autoMISST.

MD Representative Selection. Representative structures from each functionally relevant group to be used in MD simulations were selected from the autoMISST output. Representatives were chosen on the basis of the following criteria: highly significant DASP3 score, NMR structure preferred over the crystal structure, and no mutation in the active site region or to any well-described functionally important residues. Six representative structures were subjected to simulation. Two of the final nine groups did not include any known structures. One group (group 5AAA), the transcriptional regulatory group, was not the subject of simulation. Representative proteins are shown in bold in Table S2.

We performed MD simulations to better understand the conformational behavior of each of the nine classes of proteins identified by autoMISST. Starting structures for these simulations were obtained from the following data available in the RCSB PDB:[34] group 3AAA, 2KOK (chain A); group 4AA, 2MU0 (chain A); group 5BAA, 2M46 (chain A); group 6AAAAA, 2L17 (chain A); group 7AAAAAAAA, 2CD7 (chain A); group 8AAA, 2GI4 (chain A). For all simulations initiated using NMR structures (all except 2CD7), the lowest-energy conformation was selected.

Constant-pH Molecular Dynamics (CpHMD). All CpHMD simulations[41,42] utilized the ff10 force field (equivalent to ff99SB) via the leaprc.constph script. The script was also used to set the appropriate PBRadii set (mbondi2). The constant-pH input file (cpin file), which informs *sander* of the residues that should be titrated during the simulation, was made using the cpinutil.py program. Only the titratable residues (Asp, Glu, His, Lys, Tyr, and Cys) within the active site profile for each representative protein were titrated and initialized to their default states (Supplemental Methods and Table 1). For the Asp, Glu, and His residues to be marked as titratable, their identifiers were adjusted manually in the PDB file to AS4, GL4, and HIP, respectively. Explicit TIP3P water molecules were used to solvate a truncated octahedral unit cell using a 12.0 Å solvent buffer, and the systems were neutralized with ions (Na^+ or Cl^-) using *tleap*.[43] Further details regarding our CpHMD preparation and protocol can be found in the Supplemental Methods.

Unrestrained MD at pH 7 was performed in blocks of 25 ns until the structures were conformationally stable as determined by root-mean-square (RMS) comparison to their initial structure. Explicit solvent CpHMD allows a residue pair, at random, to potentially interact at a set time increment; this increment was defined to be a large value (ntcnstph = 100) due to the computationally expensive relaxation step (ntrelax = 100).[42] This set time increment for residue interaction (ntcnstph) suggests that each representative protein may need to run for a longer period to undergo the same amount of protonation state changes per residue due to the variability in the number of residues being observed in each representative [ranging from 7 to 19 residues (Supplemental Methods and Table 2)].

Representative structures were subjected to 500 ns of explicit solvent CpHMD simulation using three different random seeds. Each random seed of the simulation was run for at least 500 ns or until the trajectory demonstrated structural convergence (500−1800 ns). In the first two seeds, all titratable residues were monitored (Supplemental Methods and Table 1). In the third seed, only the titratable residues within the 10 Å active site were adjusted and monitored. No noticeable differences or abnormalities were detected among the three seeds.

Simulation Analysis. Simulations were analyzed using *cpptraj*, an AmberTools trajectory analysis tool written in C++.[44] To determine structural convergence, the backbone root-mean-square deviation (RMSD) was calculated for each trajectory using *cpptraj* and visualized using *gnuplot*.[45] Details regarding simulation analysis, including dihedral angle, pK_a, and hydrogen bond analyses, are described in the Supplemental Methods.

■ RESULTS

Functionally Relevant Clustering of ArsC Structures Identifies Two Isofunctional Clusters. The process (Figure 2) for identifying functionally relevant families that are potentially isofunctional begins with a set of ArsCs of known structures called "seed proteins" (Figure 3) and their key residues (Table S1), which represent the diversity of known structures within the superfamily. Key residues are structurally analogous among the seed proteins and are clustered at the functional site of interest [as illustrated for a representative set of ArsC-like structures (Figure 1)]. Ideally, one or more of these residues is known from experimental work to be involved in the protein's molecular function. Protein fragments that include residues within 10 Å of the key residues define the active site signature (see Methods and Figure 1). These fragments will contain both functionally relevant residues and residues that are not functionally relevant, but many will be functionally relevant. Sequence fragments within the signature are long enough for statistically relevant sequence searching.[22]

In previous work, we showed this signature definition encompasses all or most functionally relevant residues for the proteins in several superfamilies.[15,17] Throughout the results described subsequently, we demonstrate the same is true for the ArsC superfamily.

Key residues for each of the 10 ArsC seed proteins (Table S1) were input into TuLIP,[18] an iterative clustering process (described in Methods) that groups proteins with more similar active site signatures together (Figure 3A). TuLIP-identified functionally relevant groups are input into MISST.

Application of TuLIP to the key residues of the 10 seed ArsC proteins is shown in Figure 3. The first MCL clustering produced a single cluster (group 0). A DASP search of the RCSB PDB showed that this group did not self-identify (Figure 3B, blue bars on either side of the 1 × 10^{-12} DASP TuLIP score threshold[18]), which indicates the cluster is not isofunctional. MCL clustering was performed a second time, with increasing stringency of the pairwise signature score. This clustering iteration produced two groups, 1 and 2 (Figure 3A). A DASP search of the RCSB PDB using each group as input shows both are self-identifying (Figure 3C,D). Scores of the known ArsC input proteins are shown as red (group 1) and pink (group 2) bars. Those of novel ArsC proteins are shown as aqua bars. Those of LMW-PTP proteins are shown as purple bars, and those of non-ArsC proteins are shown as green bars. The group 1 DASP search identified group 2 proteins at scores less significant than the DASP threshold and vice versa, indicating a clear distinction between two functionally relevant clusters. This result is similar to that seen for enolase, peroxiredoxin, and glutathione transferase superfamilies.[18,19] (In this work, the TuLIP searches were performed with DASP; autoMISST searches were performed with DASP3. Score thresholds, different for DASP and DASP3, are described in Methods.)

Nonredundant proteins found at significant scores in groups 1 and 2 are listed in Figure 3. Five ArsCs with known structures, 2MU0, 2M46, 3RDW, 3F0I, and 3FZ4, are not clustered with either group 1 or group 2 during TuLIP searches (aqua bars, Figure 3C) and are, thus, not included as input into MISST. Similarly, several LMW-PTPs with known structures were identified at nonsignificant scores (purple bars, Figure 3D) and were not input into MISST. Singlets are known to be part of the TuLIP process, the structural database is simply not diverse enough to contain multiple members of all isofunctional groups.[18,19] All of these proteins were later identified as superfamily members during the autoMISST searches of GenBank (discussed subsequently).

We compared this TuLIP-based clustering of known protein structure with the structure-based network analysis accomplished by Atkinson and Babbitt, which identified only one ArsC group represented by structure 1I9D.[5] 1I9D is identified in group 1 of TuLIP-based structural clustering. TuLIP's structure-based clustering, which focuses on active site details, rather than on the full sequence comparison, identifies a second functionally relevant ArsC group not identified in the previously reported work.

Six Isofunctional Groups in the Arsenate Reductase Superfamily Are Identified from the First Round of autoMISST. The proteins contained in each of the two TuLIP-identified groups (Figure 3C,D) were input into autoMISST.[23] autoMISST utilizes iterative DASP3 searches[22] of GenBank to identify proteins with similar functional site features. At each iteration, groups are evaluated with respect to whether the group should be split, removed (because of significant overlap with another group), or complete (because <5% of new proteins have been identified in consecutive searches). In each search, the distribution of DASP3 scores is divided into "significant" and "nonsignificant" subsets, based on the DASP3 score threshold of 1 × 10^{-16}. In previous evaluations of DASP3 GenBank searches, proteins identified at DASP3 scores more significant than 1 × 10^{-16} are typically isofunctional with the input proteins.[32] Those sequences scoring between 1 × 10^{-8} and 1 × 10^{-16} are typically members of the superfamily, but not isofunctional with the input group. Sequences scoring as insignificant may represent a new isofunctional family within the superfamily, and the autoMISST heuristic evaluates this possibility.

In the first round of autoMISST, eight search iterations revealed six putatively isofunctional ArsC clusters (Figure 4). These six groups contain 57729 proteins (Figure 4, first round, final groups). Because more overlap than we had previously observed in other superfamilies was observed among several of the six groups (see below and Supplemental Result 1), we hypothesized that identification of functionally relevant groups was not yet complete. Therefore, the six groups were then input into the second round of autoMISST, which produced 110549 sequences in nine groups over nine search iterations. (Note that more than a year elapsed between the first round of autoMISST and the second round of autoMISST and the GenBank sequence database had grown significantly.)

To provide some understanding of how autoMISST works, a descriptive overview of how the six potentially isofunctional groups formed during the first round of autoMISST is described in Supplemental Result 2. The naming convention is described briefly in the legend of Figure 4 and more fully in Methods. Note that all search heuristics are part of the automated process.[19,23] MISST and autoMISST have been validated on multiple superfamilies;[19,23] autoMISST does not require any user intervention beyond the input proteins and key residues for each input group.

The Second Round of autoMISST Identified Nine Isofunctional Groups in the Arsenate Reductase Superfamily. Evaluation of the six groups resulting from the first round of autoMISST showed something we had not seen before: significant overlap (cross-hits) between two groups, groups 2AAAAA and 2BAAAA, at significant scores. A detailed presentation of these results can be found in Supplemental Result 1. These results suggested that a single round of autoMISST did not fully cluster this superfamily into functionally relevant groups. Therefore, the second round of autoMISST search iterations was performed.

The six groups output from the first round were used as input for the second round. The three groups that originated from TuLIP group 1 (simply termed group 1 proteins) became groups 3−5 (Figure 4, first round), while the three groups that originated from TuLIP group 2 (simply termed group 2 proteins) became groups 6−8 (Figure 4, second round). Over eight DASP3 search iterations, six groups split into nine. Specifically, group 1BABAAAA split into two groups: 5AAA and 5BAA. Group 2BAAAA split into three groups: 7AAAAAAAA, 7AAAABAA, and 7BAAAA (Figure 4, second round). At the end of the second round, 110569 ArsC sequences were identified (Figure 4); 1612 sequences could not be unambiguously assigned and were, thus, removed during the final Crosshit analysis, 1.4% of all possible identified

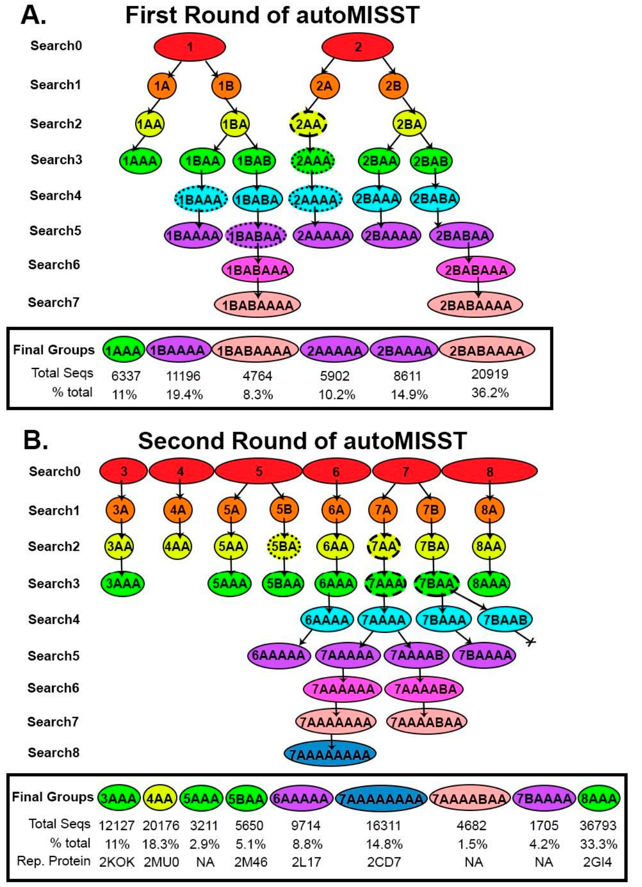


Figure 4. Progression of autoMISST search iterations shows development of nine putatively isofunctional ArsC families over two autoMISST rounds. In the first round of autoMISS (A), two input groups [the output of TuLIP (Figure 3)] resulted in six groups, each of which was used as the initial input groups for the second round of autoMISST (B), which produced nine groups. At every iteration, a group gains a letter A, unless a group is split, in which case one group gains a letter B. As described in Supplemental Methods (Combine script), sequence set removal is indicated by heavy dashed ellipse outlines (significant set) and smaller dotted ellipse outlines (insignificant set). (In this work, the "significant set" consists of those sequences in a cluster scoring better than 1×10^{-16}, while those in the "insignificant set" score between 1×10^{-8} and 1×10^{-16}. For further details, see Supplemental Methods.) For example, in round 1 of group 2AA, the significant sequence set is eliminated (heavy dash outline) because >50% of the sequences are found in another group in the third search iteration (search 2), but the insignificant set is retained and is distinctive enough to become its own functionally relevant cluster in subsequent iterations. Group 7BAAB results in a "dead end" because of autoMISST criteria for combining and removing groups (Combine script) removed both the significant and insignificant sequence sets (see Methods and Supplemental Methods).

sequences. The distribution of removed sequences in each family is found in Table S3A.

As described in the introduction, both sequence- and structure-based network analysis of the Trx-fold family identified ArsC proteins as a superfamily within the Trx-fold family.[5] A detailed review of the sequence-based similarity network created by Atkinson and Babbitt suggests that there might be three, or potentially four, families within the superfamily (see Figure 3 of ref 5). Three identifiable ArsC proteins were used in their analysis: ARSC1_ECOLX, SPX_BACSU, and YFFB_ECOLI. These three proteins are found in autoMISST groups 4AA, 5AAA, and 3AAA, respectively. Thus, the more detailed autoMISST results align with this previous work; however, the active site profile-based autoMISST process identified additional functionally relevant clusters. In the three or four families identified by sequence-based similarity networks by Atkinson and Babbitt, we could identify no group 2 proteins; however, the complete set of sequences for each of the Atkinson and Babbitt families was unavailable with the publication.

Supplemental File 1 provides lists of all sequences identified in the final DASP3 search of the second round of autoMISST, including a complete list of significantly scoring sequences (above the DASP3 score threshold of 1×10^{-16}) for each of the nine putatively isofunctional groups.

Analysis and Validation of Isofunctional Arsenate Reductase Groups Identified by autoMISST. MISST and autoMISST have been validated on several superfamilies, including the enolases and peroxiredoxins, for which significant previous work has identified isofunctional groups that allowed for robust comparison.[19,23] For the ArsC superfamily, there is no "gold standard" against which we can compare our results. Indeed, the goal of this work is not to validate autoMISST, but rather to provide new information regarding the putatively isofunctional groups within the ArsC superfamily and to identify potential mechanistic determinants within each group, thus creating potential hypotheses for future work on these proteins. However, to provide confidence in the robustness of our results, we evaluate here the distinctiveness of each group, in terms of both score distribution and cross-hit analysis.

DASP3 score distributions were analyzed first, to identify whether a distinct "trough" can be observed near a DASP3 score of 1×10^{-16} (Supplemental Result 1 and Figure 1B). The observed bimodal distribution with a trough at 1×10^{-16} is what has been observed for other protein superfamilies.[18,19] The commonality of this score threshold across superfamilies is essential: it allows the automated process, with standard heuristics for group division and completion, to be generalized across superfamilies.

Cross-hit analysis identifies protein sequences that are found at significant DASP3 scores in more than one group. Heuristics remove such cross-hit proteins from the final set of proteins (see Methods).[23] As described in Supplemental Methods, sequences are assigned to a group only if they can be unambiguously assigned. Analysis of these cross-hits illustrates the distinctiveness of each group. However, a significantly large number of cross-hits is an indication that some of the groups have not fully resolved into functionally relevant clusters.

Cross-hits were evaluated in detail for the results of the final search: the final nine groups. Unlike the results from the first round of autoMISST, the nine groups for the second round were more distinctive (Figure 5). Plotting total cross-hits versus the total number of sequences identified at each DASP3 score demonstrates a distinct inflection at the score threshold of 1×10^{-16} (Figure 5), thus providing additional support for this score threshold.

More detailed analysis of specific cross-hit numbers at each DASP3 score suggests two different types of cross-hits in the ArsC superfamily (highlighted in green and yellow in Table S3B). At the end of the first round of autoMISST, two groups, 2AAAAA and 2BAAAA, exhibited significant cross-hits. In the second round of autoMISST, four groups resulted from these two groups: 6AAAAA, 7AAAAAAAA, 7AAAABAA, and

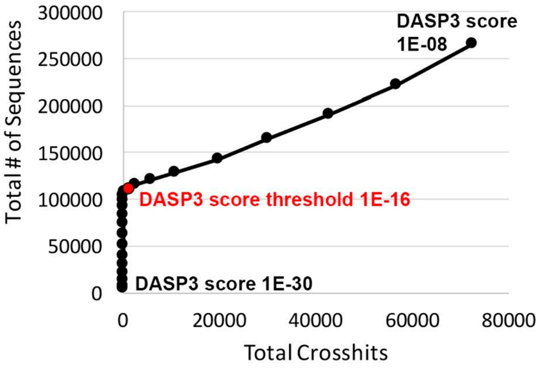


Figure 5. Cross-hit analysis of sequences identified in the final DASP3 search of the second round of autoMISST. Correlation between all identified sequences and the number of total cross-hits with changing DASP3 score. The red marker indicates the DASP3 score threshold.

7BAAAA. Two of these groups, 7AAAABAA, and 7BAAAA, are distinct, as illustrated by the observation of no cross-hits (Table S3B). The other two groups, 6AAAAA and 7AAAAAAAA, display overlap through a DASP3 score of 1 × 10^{-19} (highlighted in green in Table S3B). The number of overlapping sequences is quite small, around 300 sequences at a DASP3 score of 1 × 10^{-17}, <0.3% of all sequences in the two groups. This result suggests that, while the groups are likely functionally relevant, they are not yet fully distinguished by autoMISST or the method is not currently able to fully distinguish the functional details of these particular groups as belonging to two isofunctional groups. We suggest that the method is not yet sensitive enough to fully distinguish these particular groups.

A second and more interesting result is observed between groups 4AA and 7AAAAAAAA; a small number of sequences appeared in our cross-hit analysis. Interestingly, the DASP3 scores were very significant (as significant as 1 × 10^{-24}) in both groups (highlighted in yellow in Table S3B); this is an unusual result in our experience. These sequences were analyzed in detail (Supplemental Result 3) and found to contain two distinct active sites, one active site containing the group 4AA features and one containing the group 7AAAAAAAA features. This demonstrates that autoMISST identified two different active sites in these proteins. Most, but not all, of these sequences are from *Paracoccus* species (Supplemental Result 3 and Table 1). These results (Supplemental Result 3) suggest that two functional sites with recognizable ArsC active site signatures may have been fused into a single protein. Further evaluation of the evolution and physiological function of these 29 proteins will be necessary to determine if these multiple functional sites are physiologically or biologically relevant in these organisms.

Previously Identified Experimental Structures Are Distributed across Seven of the Nine Functionally Relevant Groups. We evaluated the groups into which ArsC proteins of known structure were placed by autoMISST (further summarized in Table S2). Initial observations indicate that most structures were annotated as ArsC, though one (2GI4) is a low-molecular weight phosphatase, two (1Z3E and 3L78) are the transcriptional regulator Spx, and one (1RW1) is annotated as YffB (Table S2).

The most well-studied ArsC reductases were identified during the first round of autoMISST. The ArsC from *Escherichia coli* plasmid R773 (1I9D), which is the widely used representative for Grx-linked ArsC proteins,[46,47] was identified as a member of group 4AA (Table S2). The best-characterized of the Trx-linked ArsC proteins[46] were identified in the second round of autoMISST, including the ArsC enzyme of *Staphylococcus aureus* plasmid pI258 (1JF8, 1JFJ, and 1LJL) and *Bacillus subtilis* ArsC (1JL3). Both were identified as members of group 7AAAAAAAA (Table S2).

The 13 proteins of known ArsC structure are distributed across seven of the nine final functionally relevant groups (Table S2). Groups 7AAAABAA and 7BAAAA are not yet represented by a crystal structure. Notably, several proteins of known structure were not input proteins: 2MU0, 2M46, 3RDW, 3F0I, and 3FZ4. Rather, these were identified during the TuLIP process at insignificant scores in group 1 and are currently annotated as ArsC proteins (aqua bars, Figure 3). Protein 2GI4 is annotated as a low-molecular weight phosphatase and was identified at an insignificant score in TuLIP group 2 along with nine other nonredundant LMW-PTPs of known structure. During autoMISST, all of these proteins were identified as members of isofunctional groups. Proteins 2MU0, 3F0I, and 3RDW are members of group 4AA (at scores of 8.44 × 10^{-19}, 5.72 × 10^{-22}, and 3.64 × 10^{-26}, respectively); 2M46 and 3FZ4 are members of group 5BAA (at scores of 1.10 × 10^{-24} and 2.68 × 10^{-23}, respectively), and 2GI4 is found in group 8AAA (and will be used subsequently in our MD studies as the representative protein for group 8AAA). This demonstrates that autoMISST can expand into functional space not represented in the input, as long as the proteins found in GenBank contain sequence fragments similar to the input profile.

One ArsC of known structure that was included in the input, 1Y1L, was not assigned to any of the nine autoMISST groups because its DASP3 score was less significant than the score threshold of 1 × 10^{-16}. This structure is unique in that it comes from an Archaea and not a bacterial species. It only shares 26% sequence identity with the family, but all three catalytic cysteines are conserved. This protein is thought to be part of the Trx-linked ArsC family, but this awaits biochemical confirmation.[46] 1Y1L is found at nonsignificant scores in all groups originating from TuLIP group 2 (Table S2), further suggesting that all functionally relevant clusters may not yet have been fully identified in this diverse superfamily.

An Introduction to Isofunctional Group Evaluation. In the next sections, we evaluate each of the nine groups, in turn, providing details about active site features, GenBank function annotations, and phylogenetic distributions (Tables 1 and 2 and Figure 7). The active site features of the groups were additionally evaluated using MD to identify potential redox mechanisms specific for each group (Figures 8−11). In all cases in which MD yielded significant insight, results are presented in respective group-specific sections.

To summarize what is described in significant detail below, ArsC groups derived from group 1 are Grx-linked ArsC, those derived from group 2 are Trx-linked or hybrid (Grx/Trx-linked) ArsC, group 3AAA contains all YffB proteins, group 5AAA consists of most transcriptional regulatory proteins (including all Spx proteins), groups 4AA and 5BAA appear to be two distinct ArsC mechanisms, and group 8AAA represents low-molecular weight protein tyrosine phosphatases.

The function annotations reported here were extracted from GenBank. Often such annotations are identified by annotation transfer between aligned sequences. As is well-documented, such annotation transfer can lead to mis-annotation when the statistical significance of the sequence alignment is not well-validated.[48−50] Once an error is made in the typical transfer approach, annotation errors are easily propagated from one protein to another.[51,52]

Annotation transfer works best for more general, rather than more specific, function annotation.[48] We have previously observed that isofunctional groups identified using active site profiling reveal more detailed biochemical function than annotations that can be identified by full sequence alignment and annotation transfer;[17−19,53] therefore, the annotations in Table 1 are a starting point based on the annotations currently in GenBank but do not represent a full description of the molecular biochemical function. We suggest that each autoMISST-identified group represents a functionally relevant and distinct cluster within the superfamily. Annotations may help provide some general insight into what those clusters may represent.

Before we analyze the MD results, the quality of the dynamics observed in our simulations was evaluated by comparison to dynamics observed in NMR structure data. Experimental and computational dynamics were found to be reasonably comparable (Figure 6). As an example, Hu and colleagues resolved and compared NMR structures of the arsenate reductase protein *Synechocystis* ArsC (SynArsC) in its four reaction stages: reduced, phosphate-bound, intermediate, and oxidized (PDB entries 2MYN, 2MYP, 2MYT, and 2MYU, respectively).[9] These proteins are similar to 2L17, the SynArsC from group 6AAAAA that we evaluated. A comparison of the backbone root-mean-square fluctuation (RMSF) plots from three seeds of 2L17 CpHMD simulations to the equivalent backbone per-residue root-mean-square deviation (RMSD) plots of the NMR structures presented by Hu and colleagues illustrates the agreement between experimental data and our simulation data. Regions of dynamic flexibility are similar, both in the NMR structures and in our MD simulations (peaks compared to bars, Figure 6). Backbone per-residue RMSD values for their four forms are identified by the red (reduced), yellow (intermediate), green (oxidized), and purple (phosphate-bound) bars in Figure 6. The reduced and intermediate forms would be most comparable to the reduced, unliganded form in our simulations. The dynamic range of ~4 Å is comparable to that observed in our simulations. These results indicate that our simulations are sampling reasonable and potentially biologically relevant protein conformations.

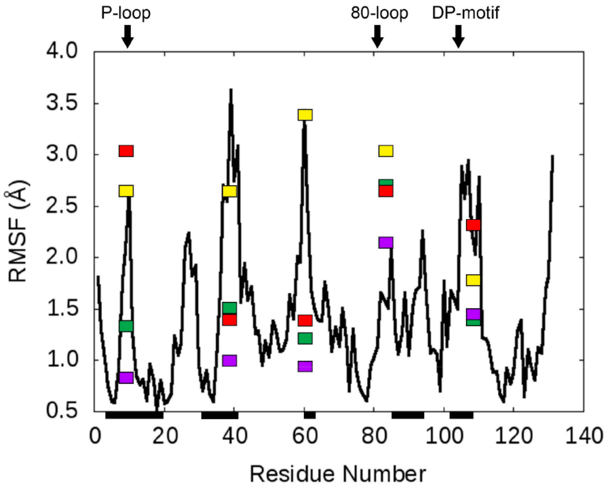


Figure 6. Evaluating our MD simulations of group 6AAAAA representative protein 2L17 in comparison to experimental data. The backbone root-mean-square fluctuation (RMSF) per residue over time during seed 1 simulation is shown in black. A higher RMSF indicates a greater deviation of said residue from its original position. The thick black lines along the *x*-axis indicate which residues correspond to fragments (Figure 1B) included in our active site signature (Figure 7B). The colored bars indicate a comparable backbone RMSF value determined by Hu and colleagues[9] (see Figure 4 in their work for exact comparison) in NMR structure experiments with the reduced (2MYN, red), phosphate-bound (2MYP, purple), intermediate (2MYT, yellow), and oxidized (2MYU, green) forms of the arsenate reductase protein, SynArsC.

Group 1 (arsenate reductase-like) and Group 2 (phosphatase-like) Proteins Display Distinct Functional Site Signatures. As described above, the first round of autoMISST identified six groups that we subsequently further divided into nine groups in the second round (Figure 4). Signature logos for the nine groups resulting from the second round of autoMISST are shown in Figure 7.

Before evaluating active site characteristics for each of the nine groups, we evaluated the common sequence motifs of the TuLIP group 1- and group 2-derived functionally relevant clusters. These motifs are observed around the active site Cys and Pro, residues highly conserved across Trx-fold family proteins.[5,36,54,55] Although it is almost invariant in all clusters, the conserved Pro was not a key residue for the second round of autoMISST, but rather the preceding residue (Arg or Asp) was used (asterisks, Figure 7). The invariant active site Cys was used as a key residue and is known to play a key role in the redox mechanism of all Trx-fold families,[2,5,36] as it also does in the ArsC proteins.

The four groups (3AAA, 4AA, 5AAA, and 5BAA) derived from TuLIP group 1 have a common cellular redox partner, Grx[46] (Table S2). Additionally, these groups share an invariant active site Cys in the Y/F-X_4-C-X_3-R/K-(R/K) motif (highlighted in gray, Figure 7A). One active site feature that distinguishes the four groups is the positively charged residue(s): KK for group 3AAA, RX for group 4AA, RK for group 5AAA, and R/K-K for group 5BAA (purple symbols, Figure 7A). This motif expands the H-X_3-C-X_3-R catalytic motif previously identified as being important for ArsC detoxification activity in the classical ArsC protein family.[56] The four N-terminal His residues with respect to the catalytic Cys are conserved in only group 4AA (green triangle, Figure 7A), a group annotated 99% as arsenate reductase (Table 1, discussed below).

Five clusters were derived from TuLIP group 2 (6AAAAA, 7AAAAAAAA, 7AAAABAA, 7BAAAA, and 8AAA), and most are Trx-linked or hybrid (Trx/Grx-linked) proteins (Table S2). The Trx-linked ArsC enzymes are similar in structure and function to the low-molecular weight protein tyrosine phosphatases (LMW-PTPs).[57,58] The TuLIP group 2-derived clusters share a Cys active site motif, N/L-F-X-C-X_2-N-X_2-R-S (highlighted in gray, Figure 7B), which suggests a connection specifically to LMW-PTPs. The classical phosphatase C-X_5-R motif[46,59] is expanded to the general PTP active site motif, C-X_5-R-S.[60,61] A conserved Asn has been observed in the X_3 position of this motif in the LMW-PTPs.[62] These comparisons indicate that the five group 2-derived groups not only are part of the ArsC superfamily but also are functionally relevant clusters within the LMW-PTP family.

Besides the redox-active Cys, a Pro, often *cis*-Pro, is another active site residue strongly conserved across the broad Trx-fold

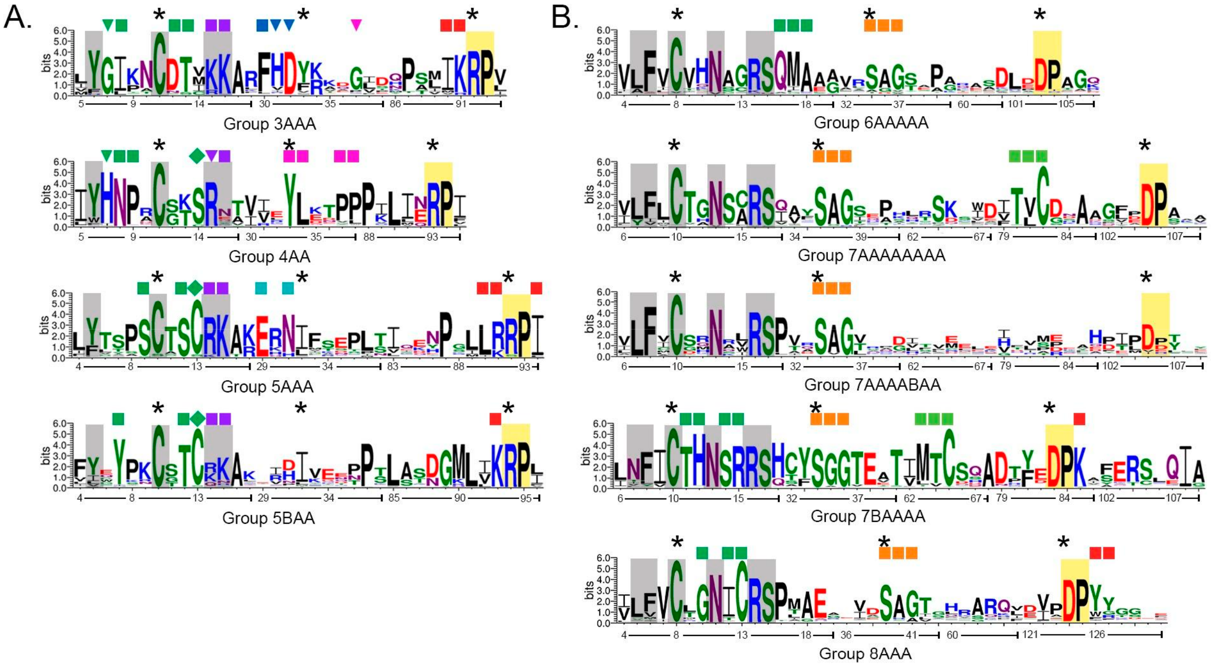


Figure 7. Signature logos for (A) group 1 and (B) group 2 show similarities and differences among the active sites of the nine ArsC groups identified by autoMISST. WebLogos were created for each cluster's active site profile with WebLogo version 3.7.[40] Each profile included active site signatures for all structures and sequences that met the DASP3 significant score threshold of 1×10^{-16}. Displayed sequences are active site signatures, as defined in Figure 1. Active site fragments (see Figure 1 for definitions) are concatenated from the N- to C-terminus. 37 and 51 comprise the signatures in clusters derived from TuLIP groups 1 and 2, respectively. Letter size signifies residue conservation within the group. The numbering within the representative protein for a group is indicated by the starting number of the fragment followed by a line. In panel A, the representative proteins used for numbering were 2KOK, 2MU0, 1Z3E, and 2M46. (The residue numbering used in this figure is based on Uniprot rather than PDB numbering. This excludes a four-residue addition in 2KOK and a 21-residue added tail in 2MU0.) In panel B, group 6AAAAA used 2L17, all groups 7 used 2CD7 numbering, and group 8AAA used 2GI4. Asterisks mark the key residues used for the second round of autoMISST. The highly conserved catalytic cysteine and proline are highlighted in gray and yellow, respectively. Other highlights and symbols, which mark conserved residues, are referenced throughout the text to highlight similarities and differences among groups.

family. This residue is located at the loop near the N-terminus of the third β strand, putting it in close juxtaposition to the redox active Cys.[5,36,54,63] This Pro is highly conserved across all but one of the autoMISST-identified groups but differs between proteins derived from TuLIP groups 1 and 2 (Figure 7). A conserved R-P motif is observed in the group 1 clusters (highlighted in yellow, Figure 7A), while a D-P motif is observed in group 2 clusters (highlighted in yellow, Figure 7B). This Pro is not as well conserved in group 7AAAABAA (highlighted in yellow, Figure 7B).

Interestingly, the Pro peptide bond configuration also distinguishes groups 1 and 2. For instance, an analysis of known structures (Table S2) reveals that the *cis* conformation is mostly observed in clusters derived from group 1 while the *trans* conformation is observed in group 2-derived clusters. The two exceptions in group 1 structures are the two proteins whose structures are determined by NMR (PDB entries 2MU0 and 2M46). No proline peptide bond isomerization was observed in our MD simulations. Mutagenesis experiments have suggested that *cis*-Pro stabilizes the protein structure in Dsb proteins[55,63] and precludes metal binding at the active site in thioredoxins.[54] Our observations that only TuLIP group 1 proteins contain active site *cis*-Pro suggest that conformation stabilization and/or prevention of metal binding may be critical for only the four groups derived from TuLIP group 1.

What is the role of the *trans*-Pro in group 2? It is interesting to observe that the group 1 and group 2 proteins correlate with Grx- and Trx-linked proteins, respectively, as well as with the *cis* and *trans* isomers of the Pro. Structural data indicate that Trx-fold proteins recognize their partner proteins by forming a transient antiparallel β sheet interaction with residues of the *cis*-Pro loop.[64] This interaction requires the *cis*-Pro loop to be able to adopt a β-strand conformation. Biophysical modeling studies have consistently shown that Asp is a nonfavored residue for β-strand structure,[65,66] so the Asp in the *trans*-Pro −1 position of the typical Trx-fold proteins would be predicted to be deleterious to this type of recognition of partner proteins.[65] The DP motif in the TuLIP group 2-derived clusters would, thus, be unfavorable to the typical Trx-fold interaction with its partner proteins.

The observation that the residue just N-terminal to the Pro is very strongly conserved as a positively charged residue (Arg) in the group 1-derived clusters and a negatively charged residue (Asp) in the group 2-derived clusters likely is mechanistically relevant. The residue just N-terminal to the Pro is known from mutagenesis studies to be very important for protein activity, with the residue identity specific to different superfamilies within the Trx-fold family.[64] Previous work on the *B. subtilis* Spx transcriptional regulator (PDB entry 1Z3E), a group 5AAA protein, found that the highly conserved R-P motif plays a role in decreasing the pK_a of the catalytic Cys.[67,68] For comparison, an Asp N-terminal to the Pro has been shown to be essential for catalysis in the dual-specificity PTPs.[69] Likewise, Bennett and co-workers indicate that this Asp is invariant in Gram-positive ArsC and very common in LMW-PTP proteins.[57] This Asp105 in *B. subtilis* ArsC (PDB entry 1JL3), a group 7AAAAAAAA protein, is proposed to serve as an acid−base catalyst, moving toward the substrate during the enzymatic reaction.[57]

The commonalities and differences identified between clusters derived from TuLIP groups 1 and 2 provide interesting fodder for understanding the origin and evolution of these proteins, which is beyond the scope of this work. In the following sections, each of the nine clusters derived by autoMISST from TuLIP clusters 1 and 2 is further distinguished by its own unique motifs and annotations. Evaluation of members of each potentially isofunctional family, as well as results of MD calculations (where relevant results were identified), can help us to understand the roles of the Arg and Asp that directly precede the active site Pro in groups 1 and 2.

A Study of Three Arsenate Reductase Active Sites: Representative Proteins from Groups 3AAA, 4AA, and 5BAA Suggest Three Distinct ArsC Mechanisms. Three groups, 3AAA, 4AA, and 5BAA, are strongly annotated as arsenate reductases, 83%, 99%, and 78%, respectively (Table 1). In our analysis discussed below, group 3AAA is represented by 2KOK, a YffB from *Brucella melitensis*,[56] and 1RW1, a YffB protein from *Pseudomonas aeruginosa*.[70] Group 4AA is represented by 1I9D and 2MU0, ArsC from *E. coli* R773[47] and *Br. melitensis*, respectively. Group 5BAA is represented by 2M46 and 3GKX, ArsC from *S. aureus* and *Bacteroides fragilis* (Table S2). Both 5BAA proteins appear to have been solved for the Structural Genomics initiative and are, thus, labeled as putative ArsC. Our work here puts both proteins firmly in a single functionally relevant cluster, group 5BAA, within the ArsC protein superfamily.

Groups 3AAA and 4AA are large families, with 12127 and 20176 proteins, respectively; group 5BAA is smaller, 5650 sequences (Figure 4). Groups 3AAA and 4AA are found largely in proteobacterial organisms, though 11% of group 4AA sequences are found in actinobacterial phyla (Table 2). More than 78% of the group 5BAA sequences are identified in firmicutes organisms (Table 2).

Observation of the overwhelming majority of ArsC annotations in these three functionally relevant clusters (Table 1) suggests three distinct ArsC functional mechanisms that are more closely related to the traditional arsenate reductase (mechanism described in Figure S1A). Potential molecular details of these three possible mechanisms are explored in the next paragraphs.

Proteins in group 4AA are annotated almost 100% as arsenate reductases (Table 1). This group is the only one that contains a conserved His residue four residues N-terminal to the redox active Cys (green triangle, Figure 7A), consistent with the H-X$_3$-C-X$_3$-R catalytic motif described as being essential for arsenate detoxification activity in what has been described as the classical arsenate reductases.[56] On this basis, we propose that the group 4AA functional site signature represents the classical ArsC mechanism.

The active site signature for group 4AA shows His8, Ser15, Arg16, and Arg94 (2MU0 residue numbering) as highly conserved residues in this family (green triangle, green diamond, purple triangle, and highlighted in yellow, respectively, Figure 7A). Mutational studies of the ArsC from plasmid R773 correlate with these results, identifying six critical amino acids: His8, Cys12, Ser15, Arg60, Arg94, and Arg107.[47,71]

We propose that the classical ArsC redox site motif represented by this group can be expanded to Y-H-N-P-X-C-X$_2$-S-R (Figure 7A). This conserved S-R motif is observed only in group 4AA (green diamond and purple triangle, Figure 7A). The conserved His has been suggested to stabilize the active site by forming a side chain H-bond to the Ser preceding the Arg in the motif in the protein structure, 1I9D.[47] On the basis of our MD simulations, expanded upon below, we propose a slightly different role for this highly conserved Ser; our observations suggest that Ser interacts with nearby residues to stabilize the N-terminal α-helical turn, thereby allowing for a helix dipole effect to aid in decreasing the pK_a of the catalytic Cys. While the Ser can also interact with the His, we believe that the Ser, by interacting with other residues, releases the His to interact with the catalytic Cys. This is critical because the His−Cys interaction is proposed to activate the Cys.[72]

Two additional conserved motifs are observed in group 4AA: Y-L-X$_2$-P-P in a second fragment (pink symbols, Figure 7A) and R-P at the active site Pro described above (highlighted in yellow, Figure 7A). In classical ArsC, three invariant arginine residues have been described: Arg60, Arg94, and Arg107 (PDB entry 1I9D residue numbering).[47] Arg94 is in the conserved Arg-*cis*-Pro motif discussed above; the other two Arg residues fall outside of the 10 Å sphere for defining active site signatures in the active site profiling method.[15]

We used MD, as described in Methods, to further explore the potential mechanistic roles played by these key residues. Structure 2MU0 was used to represent group 4AA (Table S2). Constant-pH molecular dynamics (CpHMD) simulations run over multiple trajectories revealed that the pK_a of the catalytic Cys11 was sensitive to the conformation of the residues in the active site pocket. In one trajectory in particular, when Cys11 experienced a drop in its pK_a from ∼10 to 8, the pK_a of Tyr6 and His7 simultaneously decreased and increased, respectively (Figure 8A). The range of this pK_a change is consistent with literature values.[71,73] Observations of structural changes (Figure 8B) suggest that His7 and Arg15 in 2MU0 were critical to the change in Cys pK_a, as is the N-terminal helix conformation (discussed below). In the initial structure, all three conserved residues, His7, Cys11, and Arg15, are well-separated, and Cys11 is solvent-exposed (Figure 8B, 0 ns snapshot). In the first 50−100 ns of simulation, His7 begins interacting with Arg15, and Cys11 continues to interact with solvent (Figure 8B, 107 ns snapshot). During the period between 50 and 300 ns, we see the pK_a of Cys11 stabilize at around 10 while that of His7 is 4 and that of Tyr is 12 (Figure 8A). At ∼300 ns, a conformational change occurs moving Arg15 away from His7 and to within 2.9 Å of Cys11. This effectively shields Cys11 from solvent interaction. These movements coordinate with a decrease in the Cys11 pK_a to ∼8 and an increase in the His7 pK_a to ∼6 (Figure 8A,B, 352 ns snapshot). This conformational event suggests that interactions among His7, Arg15, and Cys11—all almost invariant in this group, but not in the other groups (green and purple triangles, Figure 7A)—are critical to modulating the catalytic Cys pK_a in the classical ArsC mechanism represented by group 4AA. Perhaps these residues depict a path of electron transfer that results in the redox state change of the active site Cys.

In Trx-fold family proteins, including the ArsC proteins, the redox active Cys is in the N-terminal turn of a helix. Messens and co-workers proposed that the positive dipole effect at the N-terminal end of the helix helps decrease the Cys pK_a.[46,74] Our simulations also suggest that the helix dipole effect may

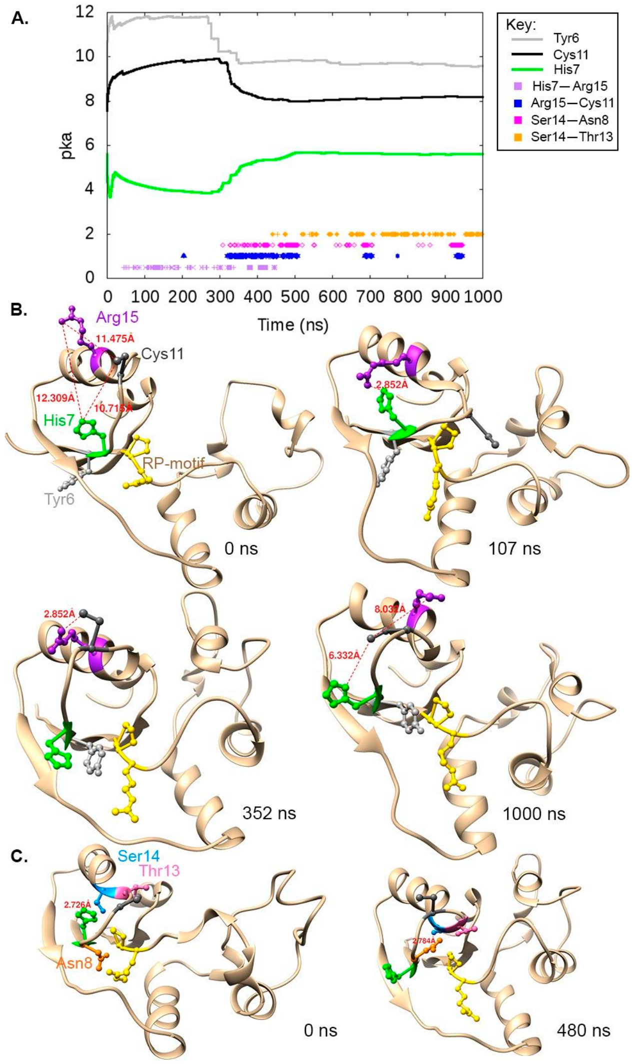


Figure 8. Side chain interactions and pK_a changes of group 4AA representative (2MU0) highlight the importance of the conserved active site His, Arg, and Ser in decreasing the pK_a of the catalytic Cys. (A) Predicted pK_a values and hydrogen bond occurrence. The simultaneous pK_a shift is a rare event seen in one of three seeds. (B and C) Snapshots illustrate protein conformation as pK_a changes. Distances are measured between the heavy atoms involved in hydrogen bonds. The R-P motif (yellow) is shown to orient viewers. (B) Four snapshots illustrate interactions among His7 (green), Cys11 (dark gray), and Arg15 (purple). (C) Two snapshots illustrate the dynamics of His7 (green) and Asn8 (orange) with Thr13 (pink) and Ser14 (blue) of the N-terminal α helix.

contribute to the Cys pK_a shift. In all group 4AA simulations, when Cys11 is solvent-exposed the pK_a remains high (10); however, the interaction with Arg described above (shown at ~300 ns in Figure 8A and at 352 ns in Figure 8B) occurs simultaneously with a tightening of the α helix, which corresponds to a decrease in the Cys11 pK_a to ~8. In all trajectories, the N-terminal helix tightening involves interactions among Ser14, highly conserved in this family (green diamond, Figure 7A), Asn9, also highly conserved in this family, (green symbol, Figure 7A), and Thr13. In the trajectory illustrated in Figure 8, the N-terminal helix tightening occurs when Ser14 hydrogen bonds with Asn8 and Thr13 (Figure 8A,C). Prior to the Cys11 pK_a shift from 10 to 8 at ~300 ns, the Ser14 side chain interacted with His7 (Figure 8A,C, 0 ns snapshot), which is not in the helix but rather is at the C-terminus of a β strand. Consistent with these observations, the His7−Ser14 hydrogen bond interactions have previously been hypothesized to be critical in stabilizing the active site.[72] This interaction with residues close to the redox Cys pulls the Cys residue into a more helical conformation at the helix N-terminus potentially contributing to the Cys pK_a shift. These results illuminate molecular functional details for group 4AA, the classical ArsC proteins.

Group 3AAA, also heavily annotated as ArsC (83%), includes some (13%) annotation as transcriptional regulator and as YffB (0.2% of proteins in this group) (Table 1), represented here by 1RW1.[70] With 3211 proteins in this family, this work significantly expands the members of the YffB-like proteins (Figure 4). The results described herein support the Buchko hypothesis of a distinction between the mechanism of classical ArsC and the YffB-like ArsC proteins and provide additional insight into the molecular characteristics that distinguish group 4AA (classical ArsC) and 3AAA (YffB-like ArsC) active sites.

Buchko and colleagues suggest that YffB proteins differ from classical ArsC, exhibiting a G-X_3-C-X_3-K motif at the redox Cys and containing a potential unique glutathione-binding site.[56] Evaluation of group 3AAA proteins expands the functional site motif proposed by Buchko, from G-X_3-C-X_3-K to Y-G-I-X_2-C-D-T-X-K-K (green symbols and highlighted in gray, Figure 7A). This differs from the active site of group 4AA (classical ArsC) in several ways, including a highly conserved Gly in group 3AAA compared to a His in group 4AA (green triangles, Figure 7A). As discussed above, this His residue is proposed to play a key function in the classical ArsC; thus, its absence in group 3AAA proteins suggests key differences in mechanism between classical ArsC (group 4AA) and YffB-like proteins (group 3AAA). In group 3AAA, the lack of a side chain in Gly (compared to the His side chain in classical ArsC) may provide additional flexibility or additional space for binding within the active site.[70] In yffB proteins, this conserved Gly residue is thought to play an important role in a potential glutathione-binding site, which is reported to be adjacent to the YffB active site motif[56] (green triangle, Figure 7A).

In addition to the active site residues near the catalytic Cys discussed above, several other residues were identified as part of the glutathione-binding site.[56] A number of these residues are conserved in the group 3AAA active site profile (blue and pink triangles, Figure 7A). The most strongly conserved residues within this proposed site are in the F-H-D motif (blue symbols, Figure 7A). Furthermore, within the same active site fragment, Gly37 is found to be moderately conserved and was proposed to be part of the proposed binding site (pink triangle, Figure 7A; 2KOK residue numbering). Two other residues, Lys35 and Glu36, were also identified as part of the potential glutathione-binding site; however, despite being identified in our analysis within the same active site fragment as the previous residues, these are not very well conserved within the active site of group 3AAA proteins (Figure 7A). We propose that the strongly conserved residues uniquely identified in the signature logo of group 3AAA that represent the potential glutathione-binding site proposed by Buchko and colleagues are, in fact, a unique component of the structure and mechanism of these YffB-like or group 3AAA ArsC proteins.

Finally, an I-K-R-P motif encompassing the active site Pro is observed in group 3AAA (red symbols and highlighted in yellow, Figure 7A). This motif is similar to that seen in group

5BAA. The Arg within this motif is almost invariant in all group 1-originating autoMISST groups and is known to be important for decreasing the pK_a of the catalytic cysteine.[46,73] Group 3AAA proteins have been shown to have only this one highly conserved Arg residue (within the R-P motif), unlike group 4AA proteins that are known to have several conserved Arg residues.[47,70] It has been hypothesized that the adjacent highly conserved Lys within this I-K-R-P motif of group 3AAA proteins may act like a conserved Arg found within the active site in the group 4AA proteins by participating in substrate binding.[70] MD simulations did not provide additional insight into the group 3AAA active site mechanism.

Group 5BAA, annotated as 78% ArsC and 18% transcriptional regulator, contains a C-X-X-C motif at the redox active site. Structural representatives, 2M46 and 3GKX from *S. aureus* and *Ba. fragilis*, respectively (Table S2), appear to have been determined for the Structural Genomics Initiative and are labeled as putative ArsC. The full redox site motif for this family is Y-X_2-C-X-T-C (green symbols and highlighted in gray, Figure 7A). As in group 3AAA, the conserved *cis*-Pro motif is a K-R-P motif (red symbols and highlighted in yellow, Figure 7A). This conserved Arg in the K-R-P motif is important for decreasing the catalytic Cys pK_a,[46,73] which is also observed in our MD simulations. During the simulations, when the conserved Arg did not interact with Cys10, the pK_a of the catalytic cysteine was found to be around 10−10.5 (Figure S2). In contrast, when the Arg did interact with the catalytic Cys, its pK_a was notably lower. Most strikingly, when a persistent interaction occurred between the Arg and Cys, a low catalytic Cys pK_a of 7.49 was observed (Figure S2). On the basis of the similarities of the *cis*-Pro motif between groups 3AAA and 5BAA, the conserved Lys in the K-R-P motif in group 5BAA likely serves a function similar to what is proposed for the conserved Lys residue in group 3AAA, i.e., substrate binding.

Group 5AAA: Heavily Annotated as a Transcriptional Regulator, Including Stress Response Regulators Spx and MgsR. Group 5AAA, also derived from TuLIP group 1, is almost entirely annotated as transcriptional regulators (98%) (Table 1). Only 1% of these sequences are labeled as ArsC, suggesting that group 5AAA represents the putatively isofunctional group of ArsC-derived transcriptional regulators. The vast majority of Spx (96%) and MgsR (38%) transcriptional regulators associated with ArsC and modulators of the general stress response are found in group 5AAA (Table 1). All group 5AAA proteins are found in bacteria, with most proteins found in the firmicutes phylum, similar to group 5BAA (Table 2).

Spx species are key regulators of the stress response, binding to the C-terminal domain of the α subunit of RNA polymerase in *Bacillus* species. Spx is conserved in other Gram-positive bacteria, where it controls the transcriptional response to oxidative stress.[67] Several known features of the Spx protein, including three conserved Arg residues, a conserved Gly52, and a ClpXP recognition sequence,[67] were not identified in the work presented here, which focused on only the ArsC-like active site and not the RNA polymerase- or ClpXP-binding sites. While the unidentified features are critical to Spx transcriptional activity, the features identified in this work represent only the ArsC redox active site. The work here suggests that group 5AAA, and the conserved ArsC functional site motifs identified within this family, represent an ArsC-like redox function in this transcriptional regulator protein family.

The CXXC motif previously recognized in the Spx proteins[67] is identified as being invariant in the active site signature of group 5AAA (Figure 7A). This is an active site feature shared with group 5BAA, but not group 3AAA or 4AA (Figure 7A). The group 1 Y/F-X_4-C-X_3-R/K-(R/K) motif is expanded to Y/F-X_4-C-X_2-C-R-K in group 5AAA (green symbols and highlighted in gray, Figure 7A). It has been proposed that, when the CXXC motif forms a disulfide bond in these proteins, Trx and Trx reductase levels are increased,[75] suggesting the mechanism by which Spx is regulated is oxidation.[67] The most distinctive motif in this group (beyond this active site Cys motif and the Arg-Pro discussed above) is the highly conserved E-X-N motif (teal symbols, Figure 7A). Both the Glu and Asn residues of this motif point toward the α helix containing the active site Cys motif. In the crystal structure (PDB entry 1Z3E), the Gln residue binds to the side chain atoms of the conserved Arg within the active site Cys motif. The role of these interactions is not known. Finally, in group 5AAA, the Arg-Pro motif is expanded to a well-conserved L-R-R-P-I motif (red symbols and highlighted in yellow, Figure 7A). MD simulations were not completed on a group 5AAA protein.

Groups 6AAAAA and 7AAAAAAAA Are Group 2 Clusters Most Heavily Annotated as Arsenate Reductase Proteins, with Some Low-Molecular Weight Tyrosine Phosphatases. Two groups of TuLIP group 2-derived proteins, groups 6AAAAA and 7AAAAAAAA, are highly annotated as ArsC, 67% and 77%, respectively (Table 1). Group 6AAAAA contains 9714 proteins, and group 7AAAAAAAA is a large potentially isofunctional family containing 16311 proteins (Figure 4). Both groups are found largely in bacteria: group 6AAAAA proteins mostly in actinobacteria and group 7AAAAAAAA proteins in proteobacterial organisms (Table 2).

Group 7AAAAAAAA has an additional unique 20% Trx annotation (Table 1) and contains the well-identified members of the Trx-linked ArsC proteins, including the ArsC from *S. aureus* plasmid pI258 (1JF8, 1JFJ, 1LJL, and 2CD7) and *B. subtilis* ArsC (1JL3 and 2IPA) (Table S2).[73,76] Group 7AAAAAAAA seems to be the well-studied Trx-linked ArsC proteins that are thought to have evolved through convergent evolution from a LMW-PTP ancestry.[9,57,58]

Group 6AAAAA contains three proteins of known structure: 3T38, 2L17, and 2L19. Both 2L17 and 2L19 are thought to be hybrid (Grx/Trx-linked) ArsC proteins, with the mechanism hypothesized for *Synechocystis* sp. PCC 6803.[77,78] This protein is typically considered to be a subgroup of Trx-linked ArsC, the hybrid ArsC, because it is structurally similar to the Trx-linked ArsC proteins but functions through a Grx-linked mechanism.[77] 3T38 is a constitutively expressed ArsC from *Corynebacterium glutamicum* and is annotated as a Trx-linked ArsC.[8]

While these proteins function as ArsC proteins, experimentally evaluated proteins in both groups show low phosphatase activity.[8,46] These Trx-linked ArsC proteins from Gram-positive bacteria can bind tetrahedral oxyanions arsenate, sulfate, and phosphate at the same ArsC active site.[57,58,74,79,80] Arsenate is the preferred substrate, with lower activities observed with sulfate and phosphate.[46]

As TuLIP group 2-derived clusters, both groups 6AAAAA and 7AAAAAAAA share the active site motif L/N-F-X-C-X_2-N-X_2-R-S as well as the conserved D-P motif (highlighted in gray and yellow, Figure 7B). They also both share a S-A/G-G motif that is largely conserved among all of the groups 2 (orange symbols, Figure 7B). The distinctive features of these groups can be seen in two places within the active site signature. In group 6AAAAA, the active site motif can be expanded by three residues: L-F-X-C-X_2-N-X_2-R-S-Q-M-A (green symbols, Figure 7B). Another highly conserved motif with a second Cys is observed in group 7AAAAAAAA: T-V-C (lime green symbols, Figure 7B). Groups 6AAAAA and 7AAAAAAAA share 0.3% of their sequences in common (Figure 5), suggesting that the ArsC site in these proteins is similar and could not be fully distinguished by autoMISST.

The mechanism of the classical Trx-linked ArsC of group 7AAAAAAAA is well-understood.[73] The catalytic mechanism requires the activity of the P-loop, which corresponds to the first fragment of our active site signature (highlighted in gray, Figure 7B), as well as an intramolecular disulfide cascade involving three conserved cysteine residues (Figure S1B).[73,74,80] In general, the mechanism of the Trx-linked ArsC involves a nucleophilic attack of the thiol of Cys10, followed by a disulfide cascade in which Cys82 attacks Cys10, then Cys89 attacks Cys82, and then Trx breaks the final disulfide bond.[74] The active site profile of group 7AAAAAAAA captures two of the three conserved Cys residues that are part of this cascade; one is within the catalytic site motif at position 10, and the other is within the T-V-C motif at position 82 (highlighted in gray and lime green symbols, Figure 7B). Mutational studies have confirmed that the first Cys residue is essential for catalysis.[7] The third conserved Cys89, while critical to overall enzymatic function, is not within our active site signature. This is likely because its role is crucial to the resetting of the ArsC enzyme through interacting with the Trx protein and not for the binding of arsenate (Figure S1B).[7,46] Keeping in mind that some proteins can have multiple active sites, and because the autoMISST methods focus on clustering proteins based on a single active site, it may not always identify important residues that are not directly related to the function of the particular active site being searched.

Although the mechanism of SynArsC has been proposed as Grx-linked, it has been shown to have three essential cysteine residues like the Trx-linked proteins, leading it to be called a hybrid ArsC.[46] Another group 6AAAAA protein, 3T38, is a hypothesized Trx-linked protein that has three catalytic Cys residues in its proposed mechanism.[8] Despite there being several Cys residues in the proposed mechanism for the members of group 6AAAAA, including both SynArsC proteins and *C. glutamicum* ArsC, group 6AAAAA does not show multiple Cys residues in its active site signature (Figure 7B). This suggests that the catalytic function of the redox active site identified in autoMISST does not require the collocation of the other two Cys residues; these Cys residues may contribute to function in another way, perhaps through the binding of a redox partner or resetting the redox switch. In SynArsC, the amino acid region around position 80 containing the latter two of the aforementioned Cys residues is in a loop conformation in its reduced state, but in a helical formation in its oxidized state.[9] This flexibility can be observed in a comparison of MD simulation to experimental NMR data in Figure 6 (see the bars under the arrow for the 80s loop). The existence of multiple conformations has been attributed to the size of the SynArsC reducing partner; a helix in the oxidized state would block the larger Trx protein and instead allow the smaller GSH to interact,[9] which further suggests that the other two active site Cys residues may contribute to the binding of the redox partner and not to the redox mechanism itself.

The representative proteins in both groups 6AAAAA and 7AAAAAAAA also have a K^+-binding pocket that can mediate efficiency but is not critical to function; it is found in the ArsC of *S. aureus* but not in ArsC from *B. subtilis*.[8,73] Thus, it follows that we did not identify high levels of residue conservation in the regions involved in the K^+-binding site around residue 62 within our active site profile (Figure 7B).

MD results for proteins from these two functionally relevant clusters derived from group 2 are described below, following our description of the annotations for the remaining group 2 clusters, so as to easily compare to results from group 8AAA.

Groups 8AAA, 7BAAAA, and 7AAAABAA Are Most Heavily Annotated as Phosphatase, Particularly Low-Molecular Weight Tyrosine Phosphatase. Three of the five putatively isofunctional clusters, 7AAAABAA, 7BAAAA, and 8AAA, derived from group 2, are heavily annotated (≥75%) as phosphatases, with very little ArsC annotation (Table 1). Groups 7AAAABAA and 7BAAAA are smaller families, with 4682 and 1705 proteins, respectively; group 8AAA is the largest group, with 36796 sequences (Figure 4).

Group 8AAA is most heavily annotated as phosphatase (93%), with a majority of the annotations as low-molecular weight (78%) and tyrosine phosphatase (81%) (Table 1). It is also the only group with a small amount of eukaryote representation [5.7% (Table 2)]. Of the sequences identified in eukaryotes, 2.76% are in metazoans and 2.08% are in fungi (Table 2). The remaining 93.8% of the proteins are found in bacteria, spread across multiple phyla, including firmicutes, bacteriodetes, actinobacteria, and proteobacteria (Table 2). Overall, of the functionally relevant clusters identified in the ArsC superfamily, group 8AAA is the most diverse, spreading across six bacterial phyla.

Groups 7AAAABAA and 7BAAAA are also annotated as phosphatase proteins (Table 1). Both of these groups include sequences but no proteins of known structure. They also have the highest percentage of hypothetical proteins at 9% and 16%, respectively (Table 1). These two findings suggest that these two clusters represent putatively isofunctional groups yet to be explored. Group 7AAAABAA is identified at moderate levels for the annotations of low-molecular weight and tyrosine phosphatase, suggesting it is a novel LMW-PTP potentially isofunctional group. Interestingly, group 7BAAAA is highly annotated as a protein tyrosine phosphatase, but not as a low-molecular weight protein (Table 1). The proteins within this group could be explored further to determine whether they make up a novel group of protein tyrosine phosphatases that share a mechanism more similar to that of the ArsC-like LMW-PTPs rather than other PTPs.

Group 8AAA includes known structures; 2G14 was used as a representative protein for this group in our MD studies. 2GI4 is annotated as a low-molecular weight phosphatase and was identified at a nonsignificant score in TuLIP group 2 (3.69×10^{-10}). Similarly, nine other nonredundant proteins were identified at nonsignificant scores in TuLIP group 2, including 4EGS, 1ZGG, 4ETI, 1U2P, 2LUO, 2CWD, 3ROF, 4ETM, and 2FEK (Table S2). All are included in group 8AAA at significant DASP3 scores. These proteins are LMW-PTPs from bacteria and eukaryotic organisms alike. Again, these results demonstrate how the diversity of the sequence database allows more complete identification of isofunctional groups within a superfamily.

Both of the human isoforms of the LMW-PTP, 1 and 2 (also called A and B, respectively), are identified in group 8AAA. Several PDBs associated with isoform 1, including 5PNT (3.46×10^{-27})[81] and 4Z99 (3.09×10^{-27}), are identified. Isoform 2 appears in group 8AAA as PDB entry 1XWW (6.49×10^{-25}).[82] These isoforms share the same redox catalytic mechanism, which relies on a single catalytic cysteine and the activity of an Arg and an Asp residue for binding (Figure S1C),[82] and this is what is captured in the active site signatures; therefore, it is expected that they would be identified within the same isofunctional family by autoMISST. The isoforms can be further distinguished by an isoenzyme-specific region between amino acids 40 and 73, likely related to their slightly different functions.[82,83] Little of this variable region is captured in the active site signatures. In the future, it would be possible to create an active site signature for each of the two isoforms, if distinguishing them as isofunctional families would be informative.

Molecular details of the group 8AAA mechanism as compared to the less understood groups 7AAAABAA and 7BAAAA are explored in the next paragraphs.

The group 2 active site motif L/N-F-X-C-X_2-N-X_2-R-S described above is expanded in group 8AAA to L-F-V-C-X-G-N—I-C-R-S-P (green symbols and highlighted in gray, Figure 7B). Differences among groups 8AAA, 7AAAABAA, and 7BAAAA are observed at positions 1, 3, 6, and 8 in this motif. This signature correlates well with the C-L-G-N-I-C-R-S phosphate-binding motif identified in the 1C0E PTP crystal structure,[62] although we do not observe conservation of the Leu at position 2 in any of these three groups. In LMW-PTPs, the second conserved Cys within the active site motif of group 8AAA (green symbols, Figure 7B) is known to be important for the regulation of protein in response to reactive oxygen, which may cause oxidation of the two Cys residues.[84] Several of the other conserved residues within this active site motif, including the active site Cys and Asn, have been identified as part of a highly conserved hydrogen bonding network in the LMW-PTPs.[85] Notably, this second Cys is not conserved in groups 7AAAABAA and 7BAAAA, suggesting a different catalytic mechanism or perhaps a different mechanism for redox regulation.

The hydrogen bonding network described in the literature includes two residues outside of the redox Cys active site motif: Ser39 and His66 (residue numbering of 2GI4).[62,86] Although His66 is not within the active site signature, Ser39 is found within the second motif distinctive to group 2: S-A/G-G (orange symbols, Figure 7B). While all other group 2 clusters share a S-A-G motif, group 7BAAAA exhibits a distinctive conserved S-G-G motif.

The third distinguishing region is the conserved D-P motif, seen in all group 2 proteins (highlighted in yellow, Figure 7B), but only weakly conserved in group 7AAAABAA. In group 7BAAAA, the motif can be expanded to D-P-K, and in group 8AAA, to D-P-Y-(Y) (red symbols and highlighted in yellow, Figure 7B). Previous studies of group 8AAA proteins have identified the conserved Y-Y motif in addition to the D-P motif (highlighted in yellow, Figure 7B) as being critical for the regulation of LMW-PTP activity through phosphorylation and dephosphorylation.[87]

Groups 7AAAABAA, 7BAAAA, and 8AAA can be further distinguished by the active site signature fragments located within the variable region, defined as amino acids 40−73 in the LMW-PTPs.[82,83] Active site signature fragments 2 and 3, which include the conserved S-A/G-G motif (orange symbols, Figure 7B), are found in this region. The variable region also includes the active site fragment that contains the M-T-C motif unique to group 7BAAAA, which is analogous to the T-V-C motif in group 7AAAAAAAA (lime green symbols, Figure 7B). In group 7AAAAAAAA, this conserved Cys is the second Cys within a trio of cysteines that comprise an intramolecular disulfide cascade.[73,74,80] The conserved Cys in the M-T-C motif in group 7BAAAA may play a role in a similar cascade. Investigating the role of this second Cys would help elucidate a catalytic mechanism for this new group.

Group 8AAA appears to be the well-described low-molecular weight phosphatase family related to ArsC. We propose that groups 7AAAABAA And 7BAAAA represent new and, as yet, unstudied LMW phosphatase mechanisms, respectively.

Comparing Active Site Details of the Group 2-Derived ArsC Proteins with LMW-PTP Proteins Identifies Distinctive Potential Pathways for Activity. Structures have been determined for three of the five putatively isofunctional clusters in the group 2-derived proteins and can, thus, be studied by MD: groups 6AAAAA and 7AAAAAAAA, both heavily annotated as ArsC, and group 8AAA, strongly annotated as phosphatase. MD simulations for each group identify distinct lower-energy active site conformations, which when taken together suggest distinct pathways for traversing high-energy backbone conformations toward more binding- or activity-competent conformations, as proposed by Karplus and colleagues.[31]

The atomistic behavior of the active site was evaluated by MD, with a key focus on the conserved Asn, which is almost invariant in all group 2-derived clusters. This residue is found within the flexible P-loop and is known to occur in both right-handed and left-handed helical conformations as part of the LMW-PTP enzyme mechanism,[88] and in either the β strand or left-hand helical conformations for ArsC proteins, depending on the protein and specific binding state.[57,62,73,74,85] The ability of Asn to assume backbone conformations that are higher-energy has been well-documented.[89] In ArsC proteins, it is hypothesized that the β-strand conformation is conserved to allow for a slightly larger binding pocket for the arsenate ligand.[57] This conformational change is particularly significant because it adjusts the distance between the active site Cys and invariant Arg[57,90,91] residue that is known to be implicated in substrate binding and to directly impact Cys pK_a.[73,74] Therefore, considering the importance of the Asn residue to both local protein conformation and active site mechanism, we clustered our MD trajectories based on the backbone conformations of the conserved Asn residue and the residue preceding it.

Group 6AAAAA, a family of 9714 proteins, annotated 67% as ArsC and 20% as phosphatase (Table 1), is evaluated first. The three determined structures in group 6AAAAA are annotated as Trx or hybrid (Trx/Grx-linked) ArsC (Table S2). We suggest that this group represents another ArsC mechanism.

MD simulations of the group 6AAAAA representative structure, 2L17, reveal that the group 6AAAAA active site was the least dynamic of the three group 2-derived proteins we evaluated. For instance, our clustering of group 6AAAAA resulted in only one significantly populated conformation relative to the same clustering in groups 7AAAAAAAA and 8AAA, which identified four and five conformations, respectively. In group 6AAAAA proteins, the residue preceding the invariant active site Asn can be His, Arg, or Gln, each a bulky residue, rather than the Gly found in several other group 2-derived functionally relevant families.

In simulations starting from three different random seeds, the highly conserved active site Asn, and the Arg immediately preceding it, sampled one major conformation (Conf1), one minor conformation (Conf2), and one very minor conformation (Conf3) (Figure 9A). In the ensemble generated across all simulations, Conf1 is represented 81% of the time, Conf2 is represented 2% of the time, and Conf3 is represented only a small fraction of a percentage of the time. The remainder of the simulation is sampling ϕ and ψ values outside of these three well-defined conformations.

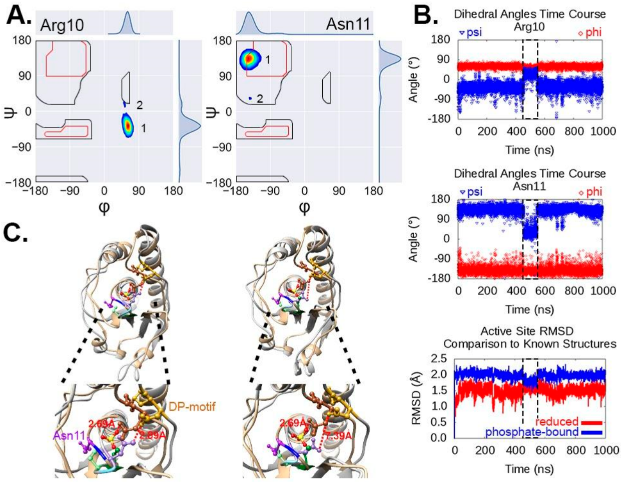


Figure 9. Conformations sampled by group 6AAAAA proteins and comparison to ligand-binding states. (A) Density plot of Arg10 and conserved Asn11. Warm colors (red and yellow) represent high density, and cooler colors (blue and green) represent low density. Arg10 Conf1 is in a high-energy region, which is discussed in the text. (B) The top two plots show the trajectory of backbone dihedral angles. The bottom plot is an RMSD comparison relative to the reduced state (PDB entry 2L17) and the phosphate-bound state (PDB entry 2L18). (C) Two conformations (gray ribbon, lighter residue color) identified in the trajectory in comparison to the phosphate-bound form of SynArsC (PDB entry 2L18), with a sulfate ion in the active site (gold ribbon, darker residue color). The binding of sulfate can be used as a first-order approximation of arsenate binding.[57] Conserved residues are colored: green for Cys, blue for Arg10, purple for Asn, and yellow for the Asp-Pro motif. The conserved D-P motif is colored to help orient the viewer. In the phosphate-bound structure (gold ribbon), the conserved Asn is in the left-handed conformation, facing away from the active site, and the conserved Asp residue interacts with the ligand.

Notably, Arg10 Conf1 is in a high-energy region, while Asn11 is in the allowed β-sheet region (Figure 9A). We asked whether this was anomalous to our MD simulations and determined that, in the NMR structures, Arg10 is also observed in high-energy conformations with a ϕ similar to what we observe in Conf1, but with a variable ψ angle.[78]

Structures of the protein members of this family are available in various forms (2L17, 2L18, and 2L19, reduced, phosphate-binding and intermediate states;[78] and 2MYN, 2MYP, 2MYT, and 2MYU, reduced, phosphate-binding, intermediate, and oxidized states[9]), so we can ask which forms are most similar to the three conformations observed in the dynamics simulations. On the basis of an RMSD comparison using the first fragment of the active site signature (residues 4−20), Conf1 is most similar to the unliganded conformation: RMSD of 1.85 Å compared to the unliganded structure and 2.17 Å compared to the phosphate-bound structure. Conf2 is most similar to the phosphate-bound form: RMSD of 1.97 Å compared to the unliganded structure and 1.83 Å compared to the phosphate-bound structure. This comparison is exemplified by a simulation time course of seed 2 (Figure 9B), though this behavior is seen in the time courses of all three simulation seeds. As shown in Figure 9B, the ψ angles of both Arg10 and Asn11 change from Conf1 (~−50° and ~135°, respectively) to Conf2 (~15° and ~35°, respectively) during the period between 400 and 600 ns, and that correlates with a noticeable change in the RMSD values relative to the reduced and phosphate-bound experimental structures. As the ψ values change, the structure becomes more similar to the phosphate-bound and less similar to the reduced structure.

Evaluation of the structures indicates that, in Conf2, the side chain of the conserved active site Asn moves farther from Asp103 (of the invariant DP motif) (Figure 9C), which begins to make space for the phosphate to bind. We thus suggest that, during the simulations, the protein is sampling a more ligand-binding-competent conformation (Conf2) a small fraction of the time. In an environment in which ligand is present, this conformation could allow the ligand to bind more easily.

The second group-2 derived functionally relevant cluster, group 7AAAAAAAA, includes two determined structures, 1JL3[76] and 1JF8, that are most often termed Trx-linked ArsC in the literature (Table S2; 2CD7, the structure used for MD simulations, is 98% similar to 1JF8; see Methods for more information about the selection of representatives). Other determined structures in this group include oxidized, reduced, disulfide intermediate, arsenite-bound, and Trx-complexed forms of these proteins (PDB entries 1JFV, 1JF8/2CD7, 1LK0, 1LJU, and 2IPA, respectively). Of the 16311 proteins in this group, 77% are annotated as arsenate reductases and 6% are annotated as phosphatases (Table 2); both representative determined structures are annotated as Trx-linked proteins (Table S2). We propose that this group represents another ArsC cluster, potentially drafted for phosphatase duty, as proposed by Messens and colleagues for structure 1JF8.[58]

The conserved active site Asn has been a focus in this group. The *S. aureus* structure (1JF8)[58] shows this Asn in the left-handed α-helical conformation when sulfate (isosteric with arsenate and phosphatase) is present (2FXI);[73] the presence of a ligand stabilizes the active site.[92] In the *B. subtilis* structure (1JL3),[57] this Asn is in the β-strand conformation when the sulfate ligand is present.[57,73]

In three randomly seeded simulations, we observed correlated motion between the active site Asn and the residue immediately preceding it, which in group 7AAAAAAAA is a conserved, but not invariant, Gly (Figure 7). In this family, His and Ala are also occasionally observed in this position (Figure 7). In group 7AAAAAAAA, the Asn and its preceding residue are found in four conformations (Figure 10), which are highly correlated. During the simulations, both residues can be found in the left-handed α-helical conformation. Gly12 is found in this conformation in Conf1 and Conf2, while Asn13 is in this conformation in Conf4. Gly12 is found in a β-strand conformation in Conf3 and at the edge of the right-handed helical conformation in Conf4. Asn13 is observed in the β-strand region for Conf1−Conf3. Across the ensemble of all

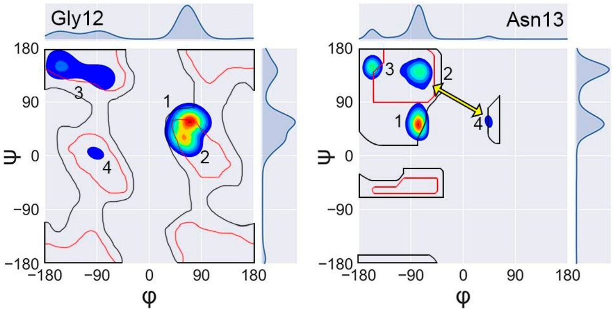


Figure 10. Evaluation of conformations sampled by group 7AAAAAAAA proteins. Density plot of the ϕ and ψ angles of conserved residue Asn13 and its preceding Gly12 residue. Warm colors (red and yellow) represent that a high density of simulation frames adopt that particular angle, and cooler colors (blue and green) represent low density. Black and red outlines correspond to approximate allowed regions computed on the basis of steric clashes between residue side chains. Outer black areas show the allowed regions if slightly shorter van der Waals radii are used in the calculation. The yellow arrow highlights the proposed transit pathway between the left-handed α helix and β-strand regions.

three simulation seeds, Conf1−Conf4 are sampled 40%, 31%, 17%, and 3% of the time, respectively.

Notably, in Conf4, the Asn adopts the left-handed helical structure for a small amount of time, similar to what is observed in the 2CD7 NMR structure in which Asn13 adopts a left-handed α helix as a minor conformation.[58] Conf4 is sampled in only one seed and is the simulation starting structure. In all simulations, the protein quickly moves away from this conformation.

What do these other conformations represent? Clearly, in these apo simulations, the Asn13 residue quickly undergoes a transition from the left-handed α helix to the β-strand conformation. At present, a conclusive link between oxyanion or ligand binding and the flip of the Asn conformation has not been resolved, but this transition appears to be essential for function.[79] However, such a transition must cross a high-energy barrier in backbone dihedral space. Brereton and Karplus proposed transit pathways, by which proteins could most easily traverse high-energy regions between low-energy conformations.[31] Correlating our MD simulation results with the Karplus pathways for traversing high-energy barriers suggests the Conf2−Conf4 pathway and vice versa (Figure 10) would be the pathway that would allow this transition. We propose this active site is designed not only for enzymatic activity but also to allow this conformational transition between two key conformations of the Asn that are required for activity.

Group 8AAA proteins are the most well-studied of the group 2-derived families. This group is annotated as 93% phosphatase, with the majority of those proteins labeled as LMW-PTP (Table 1). An array of known structures found in this family are listed in Table S2; 2GI4, a protein annotated as a LMW-PTP from *Campylobacter jejuni*,[86] was selected as the MD simulation representative.

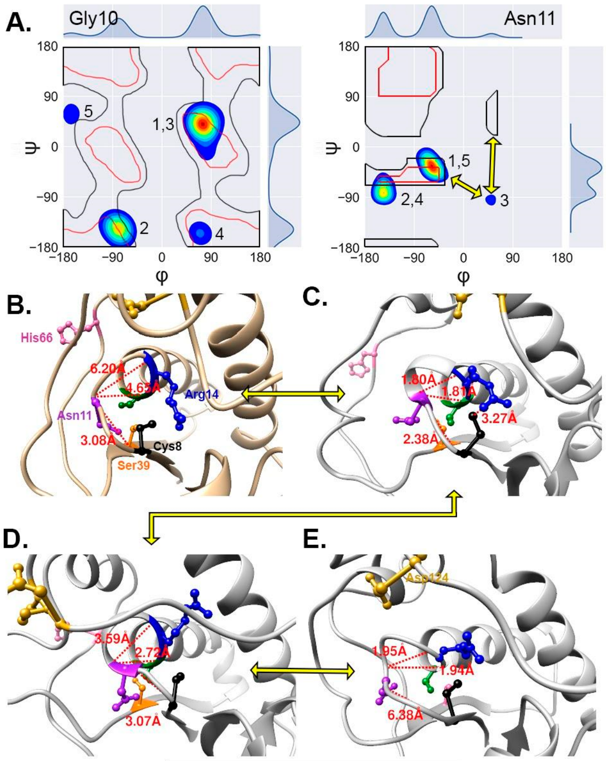
For this protein, active site profiling identified three functionally relevant protein fragments that correspond well with the residues identified as being important through experimental observation: the phosphate-binding loop (P-loop), E-loop, and D-loop. The E-loop is cited as being important to substrate specificity[86] and corresponds to the second active site fragment (Figure 7B). The P-loop, which contains the highly conserved catalytic motif, and the D-loop, which contains another highly conserved D-P motif, are important to phosphate binding and correspond to the first and fourth active site fragments, respectively (Figure 7B). In the absence of a ligand, this active site, particularly the D-loop (represented by our fourth active site fragment, Figure 7B), is flexible.[85] As with all of our simulations, apo (without a ligand present) CpHMD simulations were performed on 2GI4.

In the group 8AAA proteins, the residue immediately preceding the invariant active site Asn is Gly, and this Gly is almost invariant (Figure 7B), suggesting that the flexibility of the Gly-Asn motif may be key to the functional mechanism.[85] In most structures, the conserved Asn is observed in strained left-handed helical conformations that are stabilized by a complex hydrogen bond network involving Asn, conserved catalytic Cys, and Ser in the P-loop (asterisk and highlighted in gray, Figure 7B), the Ser in the S-A/G-G motif (orange symbols, Figure 7B), and a His that is not captured within our active site signature.[85,86,93] In several published structures, the Asn does not adopt the left-handed α helix. In these structures, the P-loop is disordered or the hydrogen bond network is disrupted and the active site residue side chains are not oriented toward the phosphate-binding site.[85,86]

In the ensemble of structures generated by three randomly seeded MD runs, Gly10 and Asn11 in 2GI4 sampled five conformations (Figure 11A), based on clustering using the ϕ and ψ angles of the conserved Gly-Asn motif. Across the ensemble, Conf1 is sampled 40% of the time, Conf2 28%, Conf3 3%, Conf4 6%, and Conf5 2%. Interestingly, while the initial structure had Asn11 in the left-handed helical conformation, the simulation quickly relaxed away from this structure, and we do not see significant sampling of this conformation again during the simulations. Conf3 is observed in a disallowed region of the Ramachandran plot in which the Asn ϕ angle (75°) is in the appropriate range for the left-handed α-helical conformation; however, the ψ angle (−90°) does not correspond to this conformation. For Conf3 (which for Gly10 overlaps Conf1), the Gly is observed mostly in the left-handed α-helical conformation, with a tiny set of conformers just outside the allowed right-handed helical conformation, similar to the Gly conformation observed in Conf4 in group 7AAAAAAAA. This rare Gly conformation outside the allowed right-handed helical conformation is sampled in 0.54% of all MD structures across the ensemble and was not further analyzed.

As described above, Brereton and Karplus have identified potential pathways that allow active site conformations to traverse high-energy areas of the Ramachandran plot, which would allow the protein active site to move to a binding-competent state, even if higher-energy conformers must be sampled.[31] On the basis of the time ordering of our simulations, the conformers identified during group 8AAA simulations map from Conf1 and Conf2 to Conf3 and then to the left-handed α-helical conformation.

Two assumptions underlie this group 8AAA active site proposal. The first is that the move from Conf3 to the ligand-binding conformation (left-handed α-helical) is a possible pathway. In the work presented here, the Conf3 Asn ϕ angle has already assumed the left-handed α-helical conformation. Only the Asn ψ angle would need to change, likely facilitated by the Gly10 conformation, for which ϕ and ψ are already in the left-handed helical conformation (Figure 11A). Consistent



Hydrogen Bonds	Conf3	Conf1	Conf2
Asn11 NH - Cys8 O	83%	7%	5%
Arg14 NH - Asn11 O	83%	2%	85%
Ser15 NH - Asn11 O	97%	91%	96%

Figure 11. Evaluation of conformations sampled by group 8AAA proteins. (A) Group 8AAA density plot of the ϕ and ψ angles of Gly10 and conserved Asn11 residues. (B–E) Snapshots show active site hydrogen bonds between conserved residues in the 2GI4 representative protein. Arrows highlight the proposed transit pathway between the left-handed and right-handed α-helical regions (B), Conf3 (C), Conf1 (D), and Conf2 (E). Conserved residues are colored: black for Cys8, purple for Asn11, blue for Arg14, green for Ser15, orange for Ser39, pink for His66, and yellow for the Asp-Pro motif. The conserved D-P motif is colored yellow to help orient the viewer. The table summarizes the percent occurrence of hydrogen bonds in each conformation.

with this proposal, the distance between Cys8 and Arg14, implicated in substrate binding[91] and Cys pK_a modulation,[57] is distinct in the three conformations and is shortest in Conf3 (average distances of 7.6 Å for Conf1, 6.5 Å for Conf2, and 4.9 Å for Conf3). This proposal is also supported by the crystal structure data of Karplus and a colleague, in which they identify this as a lower-energy pathway between the left-handed and right-handed α helix.[31]

The second assumption is that the pathway is reversible. In our simulations, which start with Asn in the left-handed helical conformation (Figure 11B), an analysis of the trajectory indicates that the pathway is sampled Conf3−Conf1−Conf2. The Conf3−Conf1 pathway is also observed by Brereton and Karplus.[31] To be biologically relevant, the reverse pathway must be possible. The Karplus data suggest that this specific high-energy pathway can be allowed in either direction.[31] Likely, our simulations were not run long enough to sample the reverse direction.

We evaluated the hydrogen bonds between conserved residues that distinguish the three conformations and may stabilize the higher-energy conformation. In Conf3, the backbone of conserved Asn11 interacted with the backbone of three of the highly conserved residues, the catalytic Cys8, Arg14, and Ser15, known to be implicated in the active site hydrogen bond network (Figure 11C, table). In Conf1, the only significantly sampled hydrogen bond was between Asn11 and Ser15 (Figure 11D, table). In Conf2, hydrogen bonds with Arg14 and Ser15 were observed (Figure 11E, table). Other hydrogen bonds involving Ser39 or His66 were not observed in our simulations, consistent with literature data[86,88] and known structures, which indicate these hydrogen bonds are transient or not present in the unliganded form.

We suggest Conf3, the highest-energy conformation of the three, according to its position on the Ramachandran plot, is a transient conformation that facilitates the transition of the Asn residue between the right-handed and left-handed α-helical conformations. The hydrogen bonds observed in Conf3 between conserved residues may help stabilize this higher-energy conformation.

During the apo simulations of the three group 2-derived putatively isofunctional groups explored here, we observe that each samples a distinct set of conformations. Comparing those dynamical conformations to known experimental structures suggests that, in the unliganded form, each active site samples conformations that represent either a more ligand-binding-competent conformation or a pathway toward a more binding-competent or active state. Our data suggest that these conformers are different for the three group 2-derived groups represented in this study of the ArsC superfamily. Specifically, the allowed pathway for group 8AAA is different from such a pathway for group 7AAAAAAAA, in which the conserved Asn does not sample Conf3 or the right-handed α-helical region. Combining our data with the results of Karplus and colleagues[31] suggests that the active site signature of each functionally relevant family may be designed, not only to perform a specific mechanism but also to distinctly traverse high-energy barriers between the lower-energy conformations that represent the unliganded state and those conformations that are more ready to bind the ligand.

■ CONCLUSION

This work has identified nine distinctly functional groups of proteins belonging to the ArsC superfamily. Five of the identified groups, groups 3AAA, 4AA, 5BAA, 6AAAAA, and 7AAAAAAAA, are predominantly annotated as ArsC reductase proteins. One group, group 5AAA, is thought to be composed of transcriptional regulators, mostly Spx proteins. Three groups, groups 7AAAABAA, 7BAAAA, and 8AAA, are highly annotated for phosphatase proteins. (It is important to recognize that sequence database annotations are often done through annotation transfer, so these annotations are only guidelines. The false annotation rate of the annotation transfer method is well-documented.[48])

Several of the autoMISST-identified functionally relevant groups are closely aligned with the groups most well-studied in the literature: the Grx-coupled "classical" ArsC (group 4AA), Trx-coupled ArsC (group 7AAAAAAAA), and LMW-PTP (group 8AAA). The other groups are less well studied. One well-known group, the Acr2 or eukaryotic ArsC, was not identified likely because its active site does not resemble that of the other known ArsC proteins.[94] Because the autoMISST approach focuses on identifying functionally relevant groups based on active site features, proteins with similar biological function but vastly different active site mechanisms are unlikely to be found in the same search. Similarly, while the LMW-PTPs were identified in this search, the other four distinctive divisions of protein tyrosine phosphatases (PTP) were not. This is likely because, despite the similarity between catalytic mechanisms among all PTPs, with classes I ("classical" PTPs), II (LMW-PTPs), and III (Tyr/Thr-specific PTPs) all relying on a cysteine-based mechanism, sequence similarity is limited to a single short active site motif, C-X-X-X-X-X-R-S/T. Furthermore, structural data comparing PTPs indicate a low degree of structural similarity, with proteins displaying different folding and topologies.[95] Conversely, the LMW-PTPs show a high degree of structural similarity with ArsC proteins, suggesting an evolutionary relationship.[57,58] Because the autoMISST process investigates the entire active site site pocket, capturing conserved components of the active sites in several fragments of the protein, and there are vast differences between LMW-PTPs and other PTPs outside of the main active site fragment, it follows that a search for the ArsC would not likely identify other PTPs.

Our work expands the knowledge of the active site of the previously and newly identified groups. We comprehensively explored the TuLIP group 1-originating groups, the Grx-coupled proteins. The active site signature of group 4AA (classical ArsC) is Y-H-N-P-X-C-X_2-S-R, while the active site signature of group 3AAA is Y-G-I-X_2-C-X-T-X-K-K. MD simulations were utilized to further explore these ArsC groups. Simulations of a classical ArsC (group 4AA) suggest the specific mechanistic importance of several conserved active site residues, including His7, Asn8, Ser14, and Arg15, to decrease the pK_a of the completely conserved catalytic cysteine. Additionally, simulations of the newly identified ArsC group, group 5BAA, reaffirm the importance of a conserved Arg residue within ArsC proteins in contributing to a lower catalytic Cys pK_a. Analysis of MD work completed on representatives of a Trx-linked ArsC group, group 7AAAAAAAA, shows a preserved right-handed conformation for the universally conserved Asn residue within the active site in the unliganded form.

Similarly, our work explored the identified phosphatase groups. All groups with phosphatase annotation originated from TuLIP group 2. They most resemble the Trx-linked ArsC proteins. Group 8AAA is well aligned with the well-studied LMW-PTP group, containing the two active isoforms of human LMW-PTPs. Simulations of this group illustrate the dynamic nature of the proteins and suggest a biologically relevant pathway for which the universally conserved active site Asn can undergo the transition from the low-energy right-handed α helix to the preserved left-handed conformation. Simulations could not be performed for the newly identified phosphatase groups because no determined structures were represented in these groups. An important next step would be to determine representatives from these functionally relevant groups.

This research opens the pathway to a more detailed biochemical understanding of ArsC and ArsC-like PTP mechanisms. Because ArsC proteins are important to arsenic redox microorganisms that decrease the concentrations of arsenate, these newly identified groups may help in the discovery of new proteins to aid in arsenate bioremediation strategies. Furthermore, the understanding of isofunctional LMW-PTP groups would support research efforts into the many diseases and disorders related to LMW-PTPs and PTPs, including cancer, neurodevelopmental disease, and diabetes.[10]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.0c00651.

> Supplemental Methods as well as supplemental results, including the rationale for two rounds of autoMISST, a detailed description of the first round of autoMISST, and a discussion of proteins with two active sites, and several tables that provide more details about the proteins identified and investigated in this work (PDF)
>
> Additional tables that provide more details about the proteins identified and investigated in this work (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Jacquelyn S. Fetrow − *Department of Chemistry, Gottwald Center for the Sciences, University of Richmond, Richmond, Virginia 23713, United States;* orcid.org/0000-0002-0528-2049; Phone: 610-921-7600; Email: jfetrow@albright.edu; Fax: 610-921-7737

Carol A. Parish − *Department of Chemistry, Gottwald Center for the Sciences, University of Richmond, Richmond, Virginia 23713, United States;* orcid.org/0000-0003-2878-3070; Phone: (804) 484-1548; Email: cparish@richmond.edu; Fax: (804) 287-1897

Authors

Mikaela R. Rosen − *Department of Chemistry, Gottwald Center for the Sciences, University of Richmond, Richmond, Virginia 23713, United States;* orcid.org/0000-0001-5941-8485

Janelle B. Leuthaeuser − *Department of Chemistry, Gottwald Center for the Sciences, University of Richmond, Richmond, Virginia 23713, United States;* orcid.org/0000-0003-1942-1422

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.0c00651

Funding

C.A.P. acknowledges support from the National Science Foundation (Grants CHE-1213271 and CHE-18800014), the Donors of the American Chemical Society Petroleum Research Fund, and the Floyd D. and Elisabeth S. Gottwald Endowment. M.R.R. acknowledges support from the University of Richmond Arts and Sciences Undergraduate Research Committee. J.B.L. and J.S.F. acknowledge research support from the University of Richmond.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Hughes, M. F. (2002) Arsenic toxicity and potential mechanisms of action. *Toxicol. Lett. 133*, 1−16.

(2) Pan, J. L., and Bardwell, J. C. (2006) The origami of thioredoxin-like folds. *Protein Sci. 15*, 2217−2227.

(3) Carvalho, A. P., Fernandes, P. A., and Ramos, M. J. (2006) Similarities and differences in the thioredoxin superfamily. *Prog. Biophys. Mol. Biol. 91*, 229−248.

(4) Lu, J., and Holmgren, A. (2014) The Thioredoxin Superfamily in Oxidative Protein Folding. *Antioxid. Redox Signaling 21*, 457−470.

(5) Atkinson, H. J., and Babbitt, P. C. (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput. Biol. 5*, No. e1000541.

(6) Fetrow, J. S., and Babbitt, P. C. (2018) New computational approaches to understanding molecular protein function. *PLoS Comput. Biol. 14*, No. e1005756.

(7) Messens, J., Hayburn, G., Desmyter, A., Laus, G., and Wyns, L. (1999) The Essential Catalytic Redox Couple in Arsenate Reductase from Staphylococcus aureus. *Biochemistry 38*, 16857−16865.

(8) Villadangos, A. F., Van Belle, K., Wahni, K., Tamu Dufe, V., Freitas, S., Nur, H., De Galan, S., Gil, J. A., Collet, J.-F., Mateos, L. M., and Messens, J. (2011) Corynebacterium glutamicum survives arsenic stress with arsenate reductases coupled to two distinct redox mechanisms. *Mol. Microbiol. 82*, 998−1014.

(9) Hu, C., Yu, C., Liu, Y., Hou, X., Liu, X., Hu, Y., and Jin, C. (2015) A Hybrid Mechanism for the Synechocystis Arsenate Reductase Revealed by Structural Snapshots during Arsenate Reduction. *J. Biol. Chem. 290*, 22262−22273.

(10) Caselli, A., Paoli, P., Santi, A., Mugnaioni, C., Toti, A., Camici, G., and Cirri, P. (2016) Low molecular weight protein tyrosine phosphatase: Multifaceted functions of an evolutionarily conserved enzyme. *Biochim. Biophys. Acta, Proteins Proteomics 1864*, 1339−1355.

(11) Zhang, Y. L., and Zhang, Z. Y. (1998) Low-affinity binding determined by titration calorimetry using a high-affinity coupling ligand: a thermodynamic study of ligand binding to protein tyrosine phosphatase 1B. *Anal. Biochem. 261*, 139−148.

(12) Lee, D., Das, S., Dawson, N. L., Dobrijevic, D., Ward, J., and Orengo, C. (2016) Novel Computational Protocols for Functionally Classifying and Characterising Serine Beta-Lactamases. *PLoS Comput. Biol. 12*, No. e1004926.

(13) Boari de Lima, E., Meira, W., and Melo-Minardi, R. C. de. (2016) Isofunctional Protein Subfamily Detection Using Data Integration and Spectral Clustering. *PLoS Comput. Biol. 12*, No. e1005001.

(14) de Melo-Minardi, R. C., Bastard, K., and Artiguenave, F. (2010) Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics 26*, 3075−3082.

(15) Cammer, S. A., Hoffman, B. T., Speir, J. A., Canady, M. A., Nelson, M. R., Knutson, S., Gallina, M., Baxter, S. M., and Fetrow, J. S. (2003) Structure-based active site profiles for genome analysis and functional family subclassification. *J. Mol. Biol. 334*, 387−401.

(16) Zhang, Y., Zagnitko, O., Rodionova, I., Osterman, A., and Godzik, A. (2011) The FGGY carbohydrate kinase family: insights into the evolution of functional specificities. *PLoS Comput. Biol. 7*, No. e1002318.

(17) Nelson, K. J., Knutson, S. T., Soito, L., Klomsiri, C., Poole, L. B., and Fetrow, J. S. (2011) Analysis of the peroxiredoxin family: using active-site structure and sequence information for global classification and residue analysis. *Proteins: Struct., Funct., Genet. 79*, 947−64.

(18) Knutson, S. T., Westwood, B. M., Leuthaeuser, J. B., Turner, B., Nguyendac, D., Shea, G., Kumar, K., Hayden, J., Harper, A., Brown, S. D., Morris, J. H., Ferrin, T. E., Babbitt, P. C., and Fetrow, J. S. (2017) An approach to functionally relevant clustering of the protein universe: Active site profile-based clustering of protein structures and sequences. *Protein Sci. 26*, 677−699.

(19) Harper, A. F., Leuthaeuser, J. B., Babbitt, P. C., Morris, J. H., Ferrin, T. E., Poole, L. B., and Fetrow, J. S. (2017) An Atlas of Peroxiredoxins Created Using an Active Site Profile-Based Approach to Functionally Relevant Clustering of Proteins. *PLoS Comput. Biol. 13*, No. e1005284.

(20) Huff, R. G. DASP. Active site profiling for identification of functional sites in protein sequences and structures. Master's Thesis, Wake Forest University, 2005.

(21) Huff, R. G., Bayram, E., Tan, H., Knutson, S. T., Knaggs, M. H., Richon, A. B.P., Santago, I., and Fetrow, J. S. (2005) Chemical and structural diversity in cyclooxygenase protein active sites. *Chem. Biodiversity 2*, 1533−1552.

(22) Leuthaeuser, J. B., Morris, J. H., Harper, A. F., Ferrin, T. E., Babbitt, P. C., and Fetrow, J. S. (2016) DASP3: identification of protein sequences belonging to functionally relevant groups. *BMC Bioinf. 17*, 458.

(23) Leuthaeuser, J. B., Hayden, J., and Fetrow, J. S. (2019) auto MISST: an automated method for functionaly relevant clustering of protein superfamilies based on active site profiling. Manuscript in preparation.

(24) Das, S., Sillitoe, I., Lee, D., Lees, J. G., Dawson, N. L., Ward, J., and Orengo, C. A. (2015) CATH FunFHMMer web server: protein functional annotations using functional family assignments. *Nucleic Acids Res. 43*, W148−153.

(25) Lee, D. A., Rentzsch, R., and Orengo, C. (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res. 38*, 720−37.

(26) Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014) Pfam: the protein families database. *Nucleic Acids Res. 42*, D222−D230.

(27) Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics 28*, 3150−3152.

(28) Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics 22*, 1658−1659.

(29) Copley, S. D. (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol. 7*, 265−272.

(30) Copley, S. D. (2012) Moonlighting is mainstream: paradigm adjustment required. *BioEssays 34*, 578−588.

(31) Brereton, A. E., and Karplus, P. A. (2015) Native proteins trap high-energy transit conformations. *Sci. Adv. 1*, No. e1501188.

(32) Leuthaeuser, J. B., Knutson, S. T., Kumar, K., Babbitt, P. C., and Fetrow, J. S. (2015) Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site microenvironment similarity. *Protein Sci. 24*, 1423−1439.

(33) Bailey, T. L., and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics 14*, 48−54.

(34) Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. (2002) The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 58*, 899−907.

(35) Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215*, 403−410.

(36) Fetrow, J. S., Godzik, A., and Skolnick, J. (1998) Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol. 282*, 703−711.

(37) Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., and Ferrin, T. E. (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinf. 12*, 436.

(38) Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res. 30*, 1575−1584.

(39) Burley, S. K., Berman, H. M., Christie, C., Duarte, J. M., Feng, Z., Westbrook, J., Young, J., and Zardecki, C. (2018) RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci. Publ. Protein Soc. 27*, 316−330.

(40) Crooks, G. E. (2004) Web Logo: A Sequence Logo Generator. *Genome Res. 14*, 1188−1190.

(41) Mongan, J., Case, D. A., and McCammon, J. A. (2004) Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem. 25*, 2038−2048.

(42) Swails, J. M., York, D. M., and Roitberg, A. E. (2014) Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using

Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput. 10*, 1341−1352.

(43) Case, D. A., Cerutti, D. S., Cheatham, T. E., III, Darden, T. A., Duke, R. E., Giese, T. J., Gohlke, H., Goetz, A. W., Homeyer, N., Izadi, S., Janawski, P., Kaus, J., Kovalenko, A., Lee, T. S., LeGrand, S., Li, P., Lin, C., Luchko, T., Luo, R., Madej, B., Mermelstein, D., Merz, K. M., Monard, G., Nguyen, H., Nguyen, H. T., Omelyan, I., Onufriev, A., Roe, D. R., Roitberg, A. E., Sagui, E., Simmerling, C. L., Botello-Smith, W. M., Swails, J. M., Walker, R. C., Wang, J., Wolf, R. M., Wu, X., Xiao, L., and Kollman, P. A. (2016) *AMBER 2016*, University of California, San Francisco, San Francisco.

(44) Roe, D. R., and Cheatham, T. E. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput. 9*, 3084−3095.

(45) Williams, T., and Kelley, C., and others (April 2013), *Gnuplot* 4.6: an interactive plotting program, http://gnuplot.sourceforge.net.

(46) Messens, J., and Silver, S. (2006) Arsenate reduction: thiol cascade chemistry with convergent evolution. *J. Mol. Biol. 362*, 1−17.

(47) Martin, P., DeMel, S., Shi, J., Gladysheva, T., Gatti, D. L., Rosen, B. P., and Edwards, B. F. (2001) Insights into the structure, solvation, and mechanism of ArsC arsenate reductase, a novel arsenic detoxification enzyme. *Structure 9*, 1071−1081.

(48) Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol. 5*, No. e1000605.

(49) Pallen, M., Wren, B., and Parkhill, J. (1999) Going wrong with confidence": misleading sequence analyses of CiaB and clpX. *Mol. Microbiol. 34*, 195.

(50) Fetrow, J. S., Siew, N., and Skolnick, J. (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J. 13*, 1866−1874.

(51) Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol. 318*, 595−608.

(52) Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci. 60*, 2637−2650.

(53) Baxter, S. M., Rosenblum, J. S., Knutson, S., Nelson, M. R., Montimurro, J. S., Di Gennaro, J. A., Speir, J. A., Burbaum, J. J., and Fetrow, J. S. (2004) Synergistic computational and experimental proteomics approaches for more accurate detection of active serine hydrolases in yeast. *Mol. Cell. Proteomics 3*, 209−225.

(54) Su, D., Berndt, C., Fomenko, D. E., Holmgren, A., and Gladyshev, V. N. (2007) A conserved cis-proline precludes metal binding by the active site thiolates in members of the thioredoxin family of proteins. *Biochemistry 46*, 6903−6910.

(55) Charbonnier, J. B., Belin, P., Moutiez, M., Stura, E. A., and Quemeneur, E. (1999) On the role of the cis-proline residue in the active site of DsbA. *Protein Sci. 8*, 96−105.

(56) Buchko, G. W., Hewitt, S. N., Napuli, A. J., Van Voorhis, W. C., and Myler, P. J. (2011) Solution structure of an arsenate reductase-related protein, YffB, from Brucella melitensis, the etiological agent responsible for brucellosis. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun. 67*, 1129−1136.

(57) Bennett, M. S., Guan, Z., Laurberg, M., and Su, X. D. (2001) Bacillus subtilis arsenate reductase is structurally and functionally similar to low molecular weight protein tyrosine phosphatases. *Proc. Natl. Acad. Sci. U. S. A. 98*, 13577−13582.

(58) Zegers, I., Martins, J. C., Willem, R., Wyns, L., and Messens, J. (2001) Arsenate reductase from S. aureus plasmid pI258 is a phosphatase drafted for redox duty. *Nat. Struct. Biol. 8*, 843−847.

(59) Mukhopadhyay, R., Zhou, Y., and Rosen, B. P. (2003) Directed evolution of a yeast arsenate reductase into a protein-tyrosine phosphatase. *J. Biol. Chem. 278*, 24476−24480.

(60) Ramponi, G., and Stefani, M. (1997) Structure and function of the low Mr phosphotyrosine protein phosphatases. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol. 1341*, 137−156.

(61) Denu, J. M., and Dixon, J. E. (1995) A catalytic mechanism for the dual-specific phosphatases. *Proc. Natl. Acad. Sci. U. S. A. 92*, 5910−5914.

(62) Tabernero, L., Evans, B. N., Tishmack, P. A., Van Etten, R. L., and Stauffacher, C. V. (1999) The structure of the bovine protein tyrosine phosphatase dimer reveals a potential self-regulation mechanism. *Biochemistry 38*, 11651−11658.

(63) Kadokura, H., Nichols, L., and Beckwith, J. (2005) Mutational alterations of the key cis proline residue that cause accumulation of enzymatic reaction intermediates of DsbA, a member of the thioredoxin superfamily. *J. Bacteriol. 187*, 1519−1522.

(64) Ren, G., Stephan, D., Xu, Z., Zheng, Y., Tang, D., Harrison, R. S., Kurz, M., Jarrott, R., Shouldice, S. R., Hiniker, A., Martin, J. L., Heras, B., and Bardwell, J. C. A. (2009) Properties of the thioredoxin fold superfamily are modulated by a single amino acid residue. *J. Biol. Chem. 284*, 10150−10159.

(65) Street, A. G., and Mayo, S. L. (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci. U. S. A. 96*, 9074−9076.

(66) Koehl, P., and Levitt, M. (1999) Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci. U. S. A. 96*, 12524−12529.

(67) Zuber, P. (2004) Spx-RNA polymerase interaction and global transcriptional control during oxidative stress. *J. Bacteriol. 186*, 1911−1918.

(68) Newberry, K. J., Nakano, S., Zuber, P., and Brennan, R. G. (2005) Crystal structure of the Bacillus subtilis anti-alpha, global transcriptional regulator, Spx, in complex with the alpha C-terminal domain of RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A. 102*, 15839−15844.

(69) Denu, J. M., Zhou, G., Guo, Y., and Dixon, J. E. (1995) The catalytic role of aspartic acid-92 in a human dual-specific protein-tyrosine-phosphatase. *Biochemistry 34*, 3396−3403.

(70) Teplyakov, A., Pullalarevu, S., Obmolova, G., Doseeva, V., Galkin, A., Herzberg, O., Dauter, M., Dauter, Z., and Gilliland, G. L. (2004) Crystal structure of the YffB protein from Pseudomonas aeruginosa suggests a glutathione-dependent thiol reductase function. *BMC Struct. Biol. 4*, 5.

(71) Gladysheva, T., Liu, J., and Rosen, B. P. (1996) His-8 lowers the pKa of the essential Cys-12 residue of the ArsC arsenate reductase of plasmid R773. *J. Biol. Chem. 271*, 33256−33260.

(72) Roos, G., Messens, J., Loverix, S., Wyns, L., and Geerlings, P. (2004) A Computational and Conceptual DFT Study on the Michaelis Complex of pI258 Arsenate Reductase. Structural Aspects and Activation of the Electrophile and Nucleophile. *J. Phys. Chem. B 108*, 17216−17225.

(73) Roos, G., Loverix, S., Brosens, E., Van Belle, K., Wyns, L., Geerlings, P., and Messens, J. (2006) The activation of electrophile, nucleophile and leaving group during the reaction catalysed by pI258 arsenate reductase. *ChemBioChem 7*, 981−989.

(74) Messens, J., Martins, J. C., Van Belle, K., Brosens, E., Desmyter, A., De Gieter, M., Wieruszeski, J.-M., Willem, R., Wyns, L., and Zegers, I. (2002) All intermediates of the arsenate reductase mechanism, including an intramolecular dynamic disulfide cascade. *Proc. Natl. Acad. Sci. U. S. A. 99*, 8506−8511.

(75) Nakano, S., Erwin, K. N., Ralle, M., and Zuber, P. (2005) Redox-sensitive transcriptional control by a thiol/disulphide switch in the global regulator, Spx. *Mol. Microbiol. 55*, 498−510.

(76) Li, Y., Hu, Y., Zhang, X., Xu, H., Lescop, E., Xia, B., and Jin, C. (2007) Conformational fluctuations coupled to the thiol-disulfide transfer between thioredoxin and arsenate reductase in Bacillus subtilis. *J. Biol. Chem. 282*, 11078−11083.

(77) Li, R., Haile, J. D., and Kennelly, P. J. (2003) An arsenate reductase from Synechocystis sp. strain PCC 6803 exhibits a novel combination of catalytic characteristics. *J. Bacteriol. 185*, 6780−6789.

(78) Yu, C., Xia, B., and Jin, C. (2011) 1H, 13C and 15N resonance assignments of the arsenate reductase from Synechocystis sp. strain PCC 6803. *Biomol. NMR Assignments 5*, 85−87.

(79) Roos, G., Buts, L., Van Belle, K., Brosens, E., Geerlings, P., Loris, R., Wyns, L., and Messens, J. (2006) Interplay between ion binding and catalysis in the thioredoxin-coupled arsenate reductase family. *J. Mol. Biol. 360*, 826−838.

(80) Zhang, W., Niu, X., Ding, J., Hu, Y., and Jin, C. (2018) Intra- and inter-protein couplings of backbone motions underlie protein thiol-disulfide exchange cascade. *Sci. Rep. 8*, 15448.

(81) Zhang, M., Stauffacher, C. V., Lin, D., and Van Etten, R. L. (1998) Crystal Structure of a Human Low Molecular Weight Phosphotyrosyl Phosphatase IMPLICATIONS FOR SUBSTRATE SPECIFICITY. *J. Biol. Chem. 273*, 21714−21720.

(82) Zabell, A. P. R., Schroff, A. D., Bain, B. E., Van Etten, R. L., Wiest, O., and Stauffacher, C. V. (2006) Crystal structure of the human B-form low molecular weight phosphotyrosyl phosphatase at 1.6-A resolution. *J. Biol. Chem. 281*, 6520−6527.

(83) Cirri, P., Fiaschi, T., Chiarugi, P., Camici, G., Manao, G., Raugei, G., and Ramponi, G. (1996) The molecular basis of the differing kinetic behavior of the two low molecular mass phosphotyrosine protein phosphatase isoforms. *J. Biol. Chem. 271*, 2604−2607.

(84) Chiarugi, P., Fiaschi, T., Taddei, M. L., Talini, D., Giannoni, E., Raugei, G., and Ramponi, G. (2001) Two Vicinal Cysteines Confer a Peculiar Redox Regulation to Low Molecular Weight Protein Tyrosine Phosphatase in Response to Platelet-derived Growth Factor Receptor Stimulation. *J. Biol. Chem. 276*, 33478−33487.

(85) Gustafson, C. L. T., Stauffacher, C. V., Hallenga, K., and Van Etten, R. L. (2005) Solution structure of the low-molecular-weight protein tyrosine phosphatase from Tritrichomonas foetus reveals a flexible phosphate binding loop. *Protein Sci. 14*, 2515−2525.

(86) Tolkatchev, D., Shaykhutdinov, R., Xu, P., Plamondon, J., Watson, D. C., Young, N. M., and Ni, F. (2006) Three-dimensional structure and ligand interactions of the low molecular weight protein tyrosine phosphatase from Campylobacter jejuni. *Protein Sci. 15*, 2381−2394.

(87) Bucciantini, M., Chiarugi, P., Cirri, P., Taddei, L., Stefani, M., Raugei, G., Nordlund, P., and Ramponi, G. (1999) The low M r phosphotyrosine protein phosphatase behaves differently when phosphorylated at Tyr131 or Tyr132 by Src kinase. *FEBS Lett. 456*, 73−78.

(88) Stehle, T., Sreeramulu, S., Löhr, F., Richter, C., Saxena, K., Jonker, H. R. A., and Schwalbe, H. (2012) The apo-structure of the low molecular weight protein-tyrosine phosphatase A (MptpA) from Mycobacterium tuberculosis allows for better target-specific drug development. *J. Biol. Chem. 287*, 34569−34582.

(89) Srinivasan, N., Anuradha, V. S., Ramakrishnan, C., Sowdhamini, R., and Balaram, P. (1994) Conformational characteristics of asparaginyl residues in proteins. *Int. J. Pept. Protein Res. 44*, 112−122.

(90) Cirri, P., Chiarugi, P., Camici, G., Manao, G., Raugei, G., Cappugi, G., and Ramponi, G. (1993) The role of Cys12, Cys17 and Arg18 in the catalytic mechanism of low-M(r) cytosolic phosphotyrosine protein phosphatase. *Eur. J. Biochem. 214*, 647−657.

(91) Madhurantakam, C., Rajakumara, E., Mazumdar, P. A., Saha, B., Mitra, D., Wiker, H. G., Sankaranarayanan, R., and Das, A. K. (2005) Crystal structure of low-molecular-weight protein tyrosine phosphatase from Mycobacterium tuberculosis at 1.9-A resolution. *J. Bacteriol. 187*, 2175−2181.

(92) Lah, N., Lah, J., Zegers, I., Wyns, L., and Messens, J. (2003) Specific potassium binding stabilizes pI258 arsenate reductase from Staphylococcus aureus. *J. Biol. Chem. 278*, 24673−24679.

(93) Evans, B., Tishmack, P. A., Pokalsky, C., Zhang, M., and Van Etten, R. L. (1996) Site-directed mutagenesis, kinetic, and spectroscopic studies of the P-loop residues in a low molecular weight protein tyrosine phosphatase. *Biochemistry 35*, 13609−13617.

(94) Fauman, E. B., Cogswell, J. P., Lovejoy, B., Rocque, W. J., Holmes, W., Montana, V. G., Piwnica-Worms, H., Rink, M. J., and Saper, M. A. (1998) Crystal Structure of the Catalytic Domain of the Human Cell Cycle Control Phosphatase, Cdc25A. *Cell 93*, 617−625.

(95) Ramponi, G., and Stefani, M. (1997) Structure and function of the low Mr phosphotyrosine protein phosphatases. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol. 1341*, 137−156.