doi:10.1016/S0022-2836(03)00056-1 J. Mol. Biol. (2003) 326, 1559–1575

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Structure and Function of an Archaeal Homolog of Survival Protein E (SurEα): An Acid Phosphatase with Purine Nucleotide Specificity

Cameron Mura[1], Jonathan E. Katz[2], Steven G. Clarke[2] and David Eisenberg[1,2]*

[1]*Howard Hughes Medical Institute and UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, 201 Boyer Hall, Box 951570, Los Angeles, CA 90095-1570, USA*

[2]*Departments of Chemistry and Biochemistry and Biological Chemistry and the Molecular Biology Institute, University of California, Los Angeles, 201 Boyer Hall, Box 951570, 611 Young Drive East, Los Angeles CA 90095, USA*

**Corresponding author*

The survival protein E (SurE) family was discovered by its correlation to stationary phase survival of *Escherichia coli* and various repair proteins involved in sustaining this and other stress-response phenotypes. In order to better understand this ancient and well-conserved protein family, we have determined the 2.0 Å resolution crystal structure of SurEα from the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum* (*Pae*). This first structure of an archaeal SurE reveals significant similarities to and differences from the only other known SurE structure, that from the eubacterium *Thermatoga maritima* (*Tma*). Both SurE monomers adopt similar folds; however, unlike the *Tma* SurE dimer, crystalline *Pae* SurEα is predominantly non-domain swapped. Comparative structural analyses of *Tma* and *Pae* SurE suggest conformationally variant regions, such as a hinge loop that may be involved in domain swapping. The putative SurE active site is highly conserved, and implies a model for SurE bound to a potential substrate, guanosine-5′-monophosphate (GMP). *Pae* SurEα has optimal acid phosphatase activity at temperatures above 90 °C, and is less specific than *Tma* SurE in terms of metal ion requirements. Substrate specificity also differs between *Pae* and *Tma* SurE, with a more specific recognition of purine nucleotides by the archaeal enzyme. Analyses of the sequences, phylogenetic distribution, and genomic organization of the SurE family reveal examples of genomes encoding multiple *surE* genes, and suggest that SurE homologs constitute a broad family of enzymes with phosphatase-like activities.

Keywords: survival protein E; domain swapping; acid phosphatase; archaeal protein; Rossmann-like fold

Introduction

The survival protein E (SurE) family was discovered nearly ten years ago by Clarke and colleagues because of its correlation to stationary phase survival of *Escherichia coli* and various repair proteins thought to be involved in creating this and other stress-response phenotypes, e.g. protein-L-isoaspartate(D-aspartate)-*O*-methyltransferase (pcm, EC 2.1.1.77).[1] The *E. coli surE* gene lies immediately upstream of the *pcm* gene, overlapping it by four nucleotides; together, these genes may form a bicistronic operon that is essential for *E. coli* viability under stressful conditions, such as elevated temperatures, osmotic stress, or high cell density. The *surE* and *pcm* genes are co-transcribed as detected by *in vitro* transcription assays,[1] although each gene may be transcribed independently from its own promoter.[2] Several bacteria contain an additional conserved gene of unknown function (*ORF0*) directly upstream of *surE*. Taken together, these several genes are thought to cluster into a stationary phase stress-survival operon, *surE-pcm-nlpD-rpoS*, where *nlpD* is an outer-membrane

Abbreviations used: SurE, survival protein E; pcm, protein-L-isoaspartate(D-aspartate)-*O*-methyltransferase; *Pae*, *Pyrobaculum aerophilum*; *Tma*, *Thermatoga maritima*; MAD, multi-wavelength anomalous dispersion; SeMet, L(+)-selenomethionine; (non-)DS, (non-) domain-swapped; RMSD, root-mean-square deviation; DDM, error-scaled difference distance matrix; ORF, open reading frame.

E-mail address of the corresponding author: david@mbi.ucla.edu

lipoprotein gene and *rpoS* encodes an alternative RNA polymerase σ factor (σ^s) that plays a regulatory role by inducing the transcription of several other stationary phase survival genes.[3]

The results reported by Visick *et al.* showed that both the *surE* and *pcm* genes are ancient and well-conserved, with orthologous genes being found in several eubacterial and archaeal species.[4] The phylogenetic distribution of *surE* genes is apparently more extensive than was initially thought, with SurE homologs having been found in eukaryotes ranging from simple protozoa (e.g. the yeast *Saccharomyces cerevisiae*) to metazoa (e.g. *Arabidopsis thaliana*). The only eubacteria in which a *surE* gene is not found are Gram-positive bacteria and mycobacteria. The results reported here emphasize the distribution and genomic organization of *surE* genes in the archaea.

The increased accumulation of isoaspartyl damage and diminished viability of stationary phase *E. coli* that have various combinations of *surE* and *pcm* mutations further supports the idea that these two proteins interact either directly or indirectly (or in parallel pathways) to provide a stress-survival phenotype: *pcm*/*surE* double mutants accumulate much higher levels of isoaspartyl residues than does the parent strain or either single mutant, and a *surE* null mutation is able to suppress stress-survival defects in a *pcm* mutant strain.[4] Recently, it was shown that the stress-survival operon noted above (*surE*···*rpoS*) was duplicated in several *E. coli* strains that were evolved over 2000 generations at high temperatures, and that these adapted *E. coli* strains displayed elevated expression from their duplicated *surE* gene compared to the ancestral lines without the duplication.[5] A conclusion of these results is that *surE* plays a significant physiological role in stress-response.

The earliest hint about the biochemical function of SurE came from genetic experiments with a protein from the yeast *Yarrowia lipolytica*. This *Y. lipolytica* protein (P30887, or *PHO2*) bears weak sequence similarity to the N-terminal domain of the SurE family, and was found to complement mutations in two of the major acid phosphatases of *S. cerevisiae*.[6] Because of its lack of sequence similarity to known phosphatases (or any other biochemically characterized protein), *PHO2* was described as a novel acid phosphatase. Most recently, the crystallographic and biochemical work of two groups has illuminated the structure and function of SurE in greater detail. Lee *et al.*[7] and Zhang *et al.*[8] independently determined the crystal structure of a SurE homolog from the eubacterium *Thermatoga maritima* (*Tma*). Their primary findings were that: (i) the *Tma* SurE monomer consists of an N-terminal globular domain of about 180 residues that resembles a Rossmann fold and a novel, extended C-terminal region of about 70 residues; (ii) monomers assemble into dimers (and possibly tetramers) with extended C-terminal α-helices that domain swap; (iii) *Tma* SurE exhibits a divalent cation-dependent acid phosphatase activity that is inhibited by vanadate or tungstate; (iv) divalent metal ions bind in a putative conserved active site; and (v) the *Tma* enzyme shows no protease or nuclease activity, but has a slight preference for guanosine 5′-monophosphate (GMP) or adenosine 5′-monophosphate (AMP) substrates. These results suggested for the first time that SurE may be a novel acid phosphatase.

In order to better understand the structures and functions of members of this ancient and well-conserved protein family, we determined the 2.0 Å-resolution crystal structure of SurEα from the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum* (*Pae*; t_{max} 104 °C).[9] Comparison to the eubacterial *Tma* SurE structure reveals several significant similarities and differences between these SurEs (such comparisons are justified because the *Pae* structure was determined independently of *Tma*, by multiwavelength anomalous dispersion (MAD) phasing). Results are presented from biochemical characterization of the acid phosphatase activities of *Pae* and *Tma* SurEs, and an analysis of the phylogenetic distribution and genomic organization of *surE* genes is given. These results are discussed in terms of the *Pae* SurEα structure, its likely biochemical function, and the SurE protein family in general.

Results

Structure determination, refinement, and validation

The *Pae* SurEα crystal structure was determined to 2.0 Å-resolution by MAD phasing of data from a single SeMet-substituted crystal, since molecular replacement efforts with *Tma* SurE homology models were unsuccessful. Despite poor phasing statistics (Table 1), electron density maps calculated with experimental MAD phases were of excellent quality (Figure 1(d) of the Supplementary Material). The high-resolution limit of the data (2.0 Å) at the SeMet peak wavelength permitted automatic model building for much of the structure by successive steps of secondary structure fragment matching (MAID) and free-atom model refinement (wARP). The final model was refined against the best dataset from the three-wavelength MAD experiment, since native, underivatized crystals were never obtained (as described in Materials and Methods). Selenium occupancies were refined in the final model and non-crystallographic symmetry restraints were not imposed after the first few rounds of model refinement. The final model contains a dimer in the asymmetric unit, with 276 of 280 residues for one monomer and 278 of 280 for the other. A total of 287 water molecules, seven glycerol molecules, and two acetate groups were modeled as solvent. Final R/R_{free} values are 18.5%/22.3%, with acceptable model geometry

Table 1. Crystallographic statistics for *Pae* SurEα

	Data collection and MAD phasing			Model refinement	
Data set	Inflection	Peak	High-energy remote	Resolution range (Å)	20.0–2.0
Wavelength (Å)	0.97870	0.97860	0.96485	No. reflections (working/test set)	38,377/2033
Resolution range (Å)	100.0–2.40	100.0–2.00	100.0–2.85	No. protein residues (A/B)[a]	276/278
No. reflections (total/unique)	242,223/ 47,599	297,069/ 42,129	206,830/ 28,846	$\langle B \rangle$ (protein atoms, Å^2)	35.73
Completeness (%)[b]	100.0 [100.0]	99.9 [100.0]	99.9 [100.0]	$\langle B \rangle$ (Wilson plot, Å^2)	28.79
$I/\sigma(I)$	15.9 [2.3]	17.4 [2.5]	14.0 [3.3]	*No. solvent molecules* ($\langle B \rangle$, Å^2)	
R_{merge} (%)[c]	9.8 [77.8]	9.9 [78.1]	14.7 [94.1]	Water	287 (44.8)
Anomalous signal ($\langle \chi^2 \rangle$)[d]	–	4.3 [1.9]	–	Glycerol	7 (73.9)
No. Se sites per a.u. (used/expected)[e]	–	8/12	–	Acetate	2 (59.8)
Phasing resolution range (Å)	38.0–2.85	38.0–2.85	38.0–2.85	No. Ramachandran violations	2/560 residues
R_{cullis}[f] Acentric	–	0.94/0.65	0.96/0.95	RMSDs (bonds (Å)/angles (deg.))	18.5/22.3
Centric	–	0.90	0.93	R_{cryst}/R_{free} (%)[g]	0.014/1.79
Figure of merit[h]	–	0.43/0.59	–	PDB submission code	1L5X

[a] Number of SurEα residues built in monomers A and B, out of 280 residues per monomer of recombinant protein (the His-tag and linker add 14 residues to the wild-type sequence).
[b] Statistics for the highest-resolution shell are given in square brackets.
[c] $R_{merge}(I) = \Sigma_{hkl}((\Sigma_i|I_{hkl,i} - \langle I_{hkl} \rangle|)/\Sigma_i I_{hkl,i})$.
[d] Anomalous signal as measured by the normalized χ^2 for merging Bijvoet pairs I^+, I^-. That is, $\chi^2 = \Sigma_{I+,I-}((I - \langle I \rangle)^2/\sigma^2(n/n-1))$. Values >2 suggest a usefully strong anomalous signal.
[e] Number of Se sites calculated by SHELXD and used for phasing (out of 12 sites expected per a.u.).
[f] $R_{cullis} = (\Sigma_{hkl}||F_{PH} \pm |F_P| - F_{H,calc}||)/\Sigma_{hkl}|F_{PH} \pm F_P|$. Statistics for acentric reflections are given as isomorphous/anomalous.
[g] $R_{cryst} = \Sigma_{hkl}||F_{obs}| - |F_{calc}||/\Sigma_{hkl}|F_{obs}|$. R_{free} was computed identically, except that 4.6% of the reflections were omitted as a test set.
[h] Values are given before/after density modification and phase extension to 2.0 Å.

and root-mean-square deviations (RMSD) from ideal values (Table 1). Some results from model validation, along with a representative example of the agreement between the final model and experimental and $2F_o - F_c$ electron density maps are shown in Figure 1 of the Supplementary Material. The Ramachandran plot shows a single residue (Ser99) in a disfavored/high-energy region of $\phi-\psi$ space (the electron densities illustrated in Figure 1 of the Supplementary Material show that this residue is modeled correctly.)

The topology and fold of SurEs

Like *Tma* SurE, *Pae* SurEα adopts a Rossmann-like fold with an extended C-terminal domain. The secondary structure elements of *Pae* SurEα are mapped to a multiple sequence alignment of archaeal SurEs in Figure 1, and the topology of this mixed α/β protein is shown schematically in Figure 2. The N-terminal core domain of ≈170 residues is a Rossmann-like fold, while the C-terminal region of ≈90 residues (gray-shaded background in Figures 1 and 2) forms an irregular structure that is dominated by a 40 residue β-hairpin. This hairpin protrudes from the body of the protein and mediates possible tetramerization of both *Pae* and *Tma* SurE (discussed below). Despite their Rossmann folds, no structure in the PDB is significantly similar to *Tma* or *Pae* SurEα. Three structural homology searches were performed with the DALI program: using the entire SurEα monomer, using the highly conserved N-terminal domain alone, or using the C-terminal region alone. The closest match was the N-terminal domain of SurEα against the Rossmann fold of phosphofructokinase, but this has a *Z*-score of only 6. Similar results were found by Lee *et al.* and Zhang *et al.* for *Tma* SurE.[7,8] As has been observed for other nucleotide-binding proteins that utilize Rossmann folds, the most highly conserved residues in the SurE family (Figure 1) map to the C-terminal loops of the β-strands that form the core of this fold (Figure 2). Two of the largest structural differences between *Tma* and *Pae* SurEα are indicated in Figure 2: (i) the more extended β-sheet core of *Tma* SurE contains seven strands rather than five, and (ii) the C-terminal α-helices exchange in *Tma* to form a domain-swapped dimer.

Comparative structural analysis of SurE monomers

In order to more quantitatively dissect the structural differences between the *Tma* and *Pae* structures, we utilized error-scaled difference distance matrices (DDMs). This is a recent structure comparison approach that explicitly takes into account the crystallographic data for the two models under comparison (e.g. resolution, *B*-factors, R_{free}) *via* a diffraction precision index.[10] The output from a pairwise comparison of models is a DDM, which

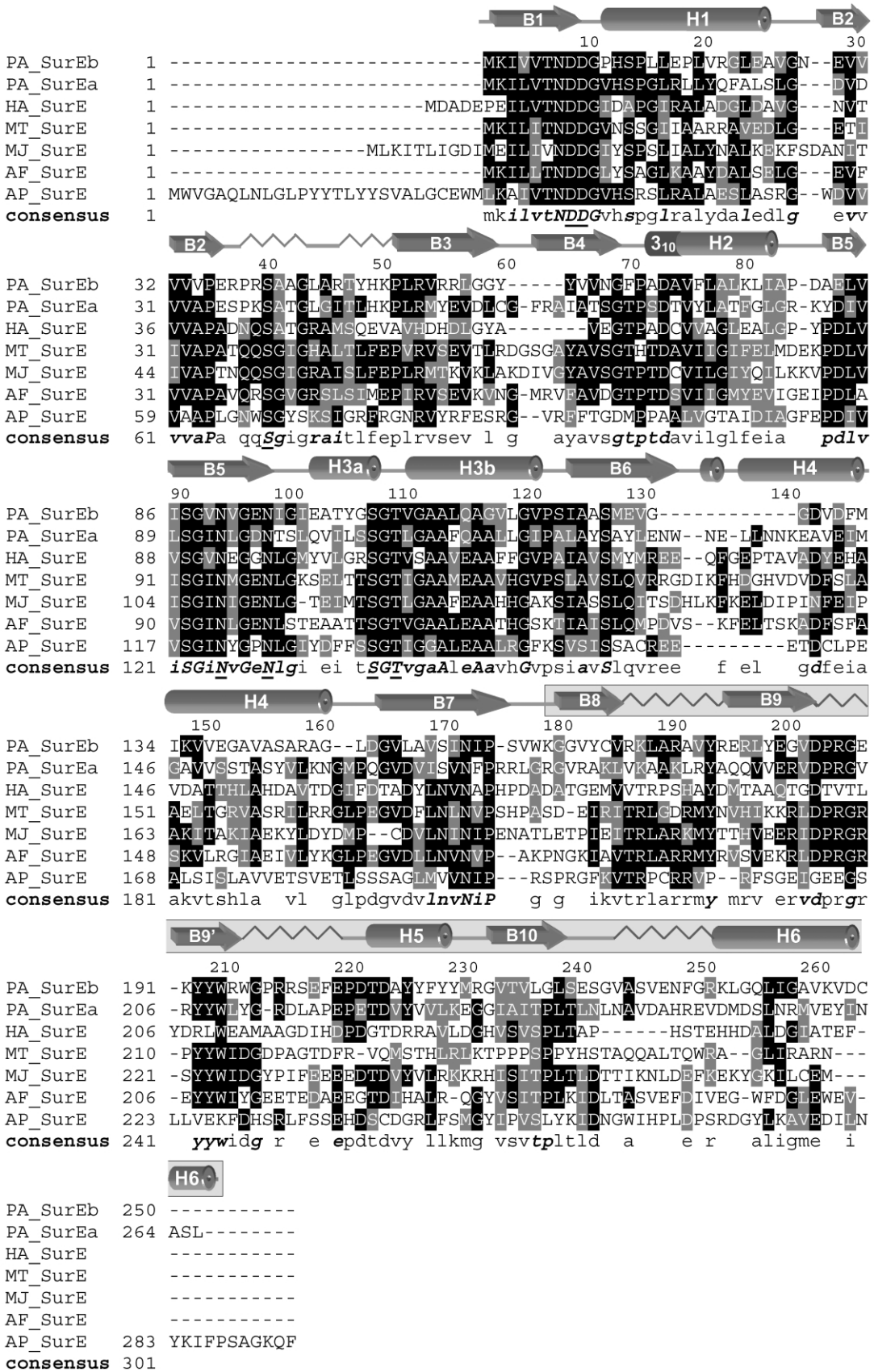

Figure 1. Sequence analysis and secondary structure of archaeal SurEs. A multiple sequence alignment is shown for all known archaeal SurE sequences. Clusters of conserved residues are shaded in black (stringent) or gray (less stringent), and the consensus sequence is given in the last line (conservative substitutions are italicized, and identities are capitalized). Numbering is for the *Pae* SurEα sequence. Regular secondary structure elements are shown as arrows

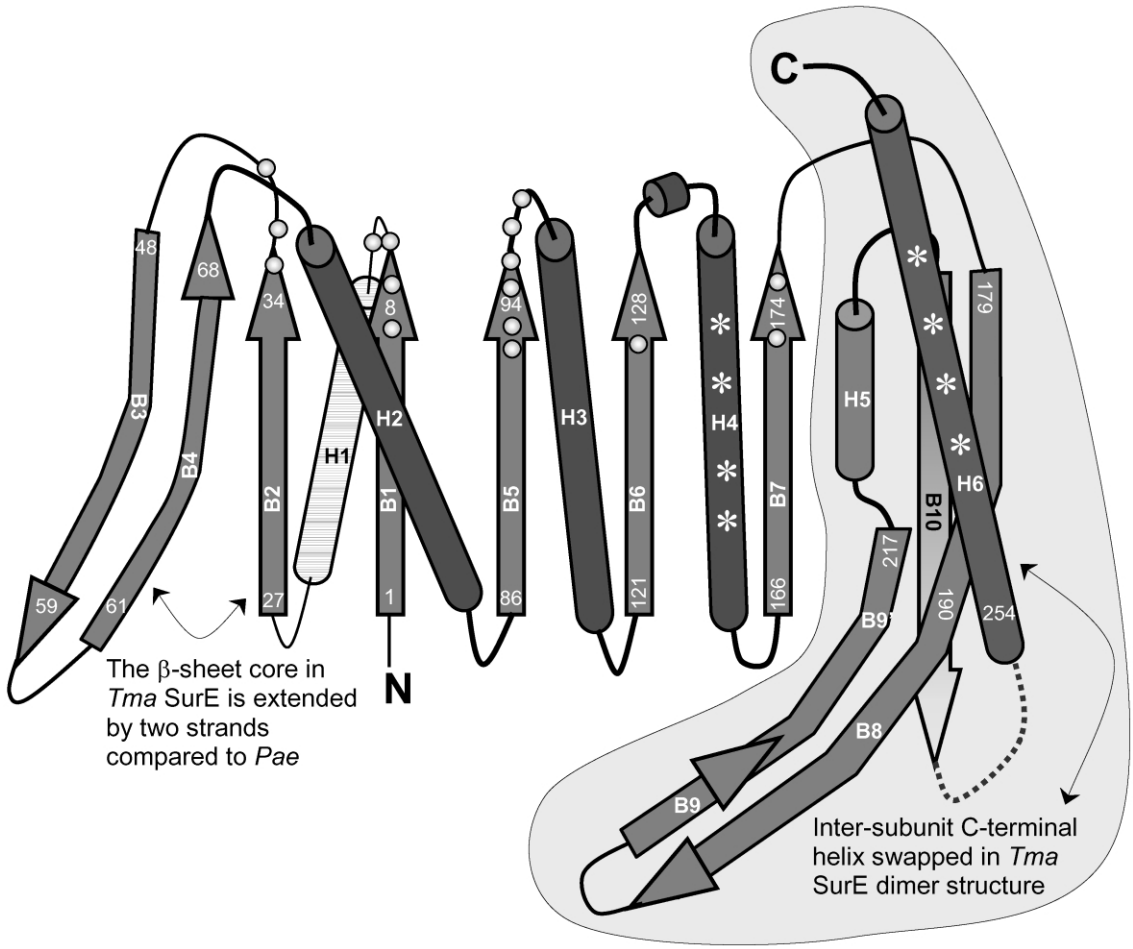

Figure 2. *Pae* SurEα is a Rossmann-like fold with an extended C-terminal domain. This cartoon of the *Pae* SurEα monomer topology illustrates the N-terminal Rossmann fold and extended C-terminal region (gray-shaded background). Residue numbers indicate the termini of secondary structural elements, and the intensity of shading conveys approximate 3D structure: lighter secondary structural elements are below the plane of the paper and darker elements are above it (asterisks on H4 and H6 denote positions that are near in 3D space). The most conserved residues in the SurE family (circles) map to the C-terminal loops of the β-strands that form the Rossmann fold. The primary differences between the *Pae* and *Tma* SurE structures (double-headed arrows) and the hinge loop region (broken line) are indicated.

is a symmetric matrix whose elements are the statistical significance of the deviation from the mean of the two structures (expressed as a standard deviation). We used a genetic algorithm to simultaneously compare all pairwise DDMs of the *Tma* and *Pae* SurEα structures,[11] examining: (i) the ensemble of eight crystallographically-independent *Tma* SurE models (two monomers per asymmetric unit in each of the PDB entries 1J9J, 1J9K, 1J9L, 1ILV) and (ii) various *Tma-Pae* SurEα pairs.

Conformationally variant regions in SurEs were revealed by application of difference distance matrices to *Pae-Pae*, *Pae-Tma*, and *Tma-Tma* SurE structure alignments. The results of these structural comparisons, as well as a standard 3D structure superposition, are shown in Figure 3. The superposition in Figure 3(a) shows that the N-terminal β-sheet core is highly conserved (1.1 Å RMSD over all atoms). Minor structural differences occur in the β-turn between strands B3 and B4 (green arrow) and the N-terminal region of the interrupted helix H4 (purple arrow). The largest rigid-body differences occur in the C-terminal β-hairpin (orange arrow, Figure 3(a)) and C-terminal α-helix, which is either swapped (*Tma*) or mostly non-swapped (*Pae*) depending on the hinge loop (red arrow). For an ensemble of eight structures, there are 28 unique difference distance matrices. An

and cylinders, and pleated lines indicate regions that can be classified only loosely as β-strands or turns. Secondary structures that form the irregular C-terminal region are shaded in a gray box (see Figure 2). The strictest conservation occurs in the N-terminal ≈150-residue core, which is also where all of the putative active site residues are located (underlined).

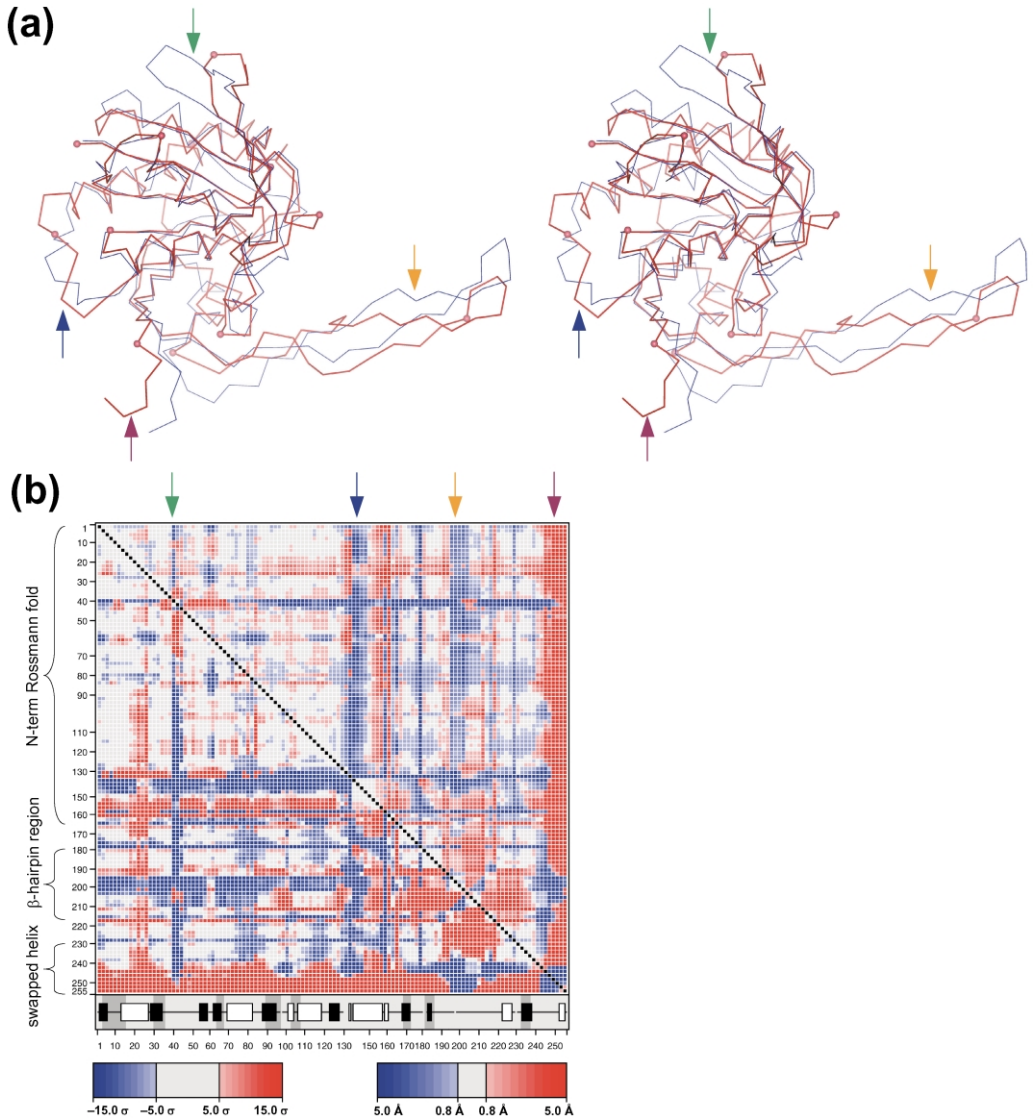

Figure 3. Structural comparison of *Pae* and *Tma* SurE reveals conformationally invariant regions. (a) A stereoview is shown of *Pae* SurEα (red) superimposed on a C^{α} trace of the *Tma* SurE monomer (blue). Swapped (*Tma*) or unswapped (*Pae*) C-terminal helices are not shown, for clarity, and a red ball marks every 20th *Pae* residue. Major structural differences are indicated by colored arrows and discussed in the text. These differences are quantified in (b) *via* error-scaled difference distance matrices.[11] An example of the pairwise comparisons between matching fragments of *Tma* and *Pae* SurEα monomers is shown in the lower triangle, and the upper triangle of (b) provides a normal matrix of RMSDs between residues in the aligned *Pae*-*Tma* structures. Color intensity scales with either the statistical significance of the model differences (below diagonal) or the actual values of differences in distances between the atomic coordinates (above diagonal).

example of one of these DDMs between *Tma* and *Pae* SurEα is shown in Figure 3(b). The lower triangle gives the statistical significance of the scaled structural differences and the upper triangle shows the actual differences between atom pairs in terms of 3D distances. Regions of relatively minor structural variation between *Pae* and *Tma* are apparent from the DDM analysis (see arrows in Figure 3(b) and corresponding differences in Figure 3(a)). For example, β-strand B4 hydrogen bonds to the adjacent strand (B2) in *Tma* to extend the β-sheet core, whereas in *Pae* SurEα this strand curves away from the sheet (see Figure 2 and green arrows in Figure 3).

Refinement of the hinge loop and domain swapping

Crystalline *Pae* SurEα is probably a mixture of domain-swapped (DS) and non-domain-swapped (non-DS) states. Residues 244–248 in *Pae* SurEα (Ala_{244}ValAspAlaHis$_{248}$) correspond to the putative hinge loop segment in the DS *Tma* SurE structures[7,8] (these residues are not conserved between *Pae* and *Tma* or among archaeal SurEs; see Figure 1). Electron density for this region is poorer than in any other part of *Pae* SurEα (Figure 4). In order to distinguish the DS from the non-DS conformation of the hinge loop region, occupancies

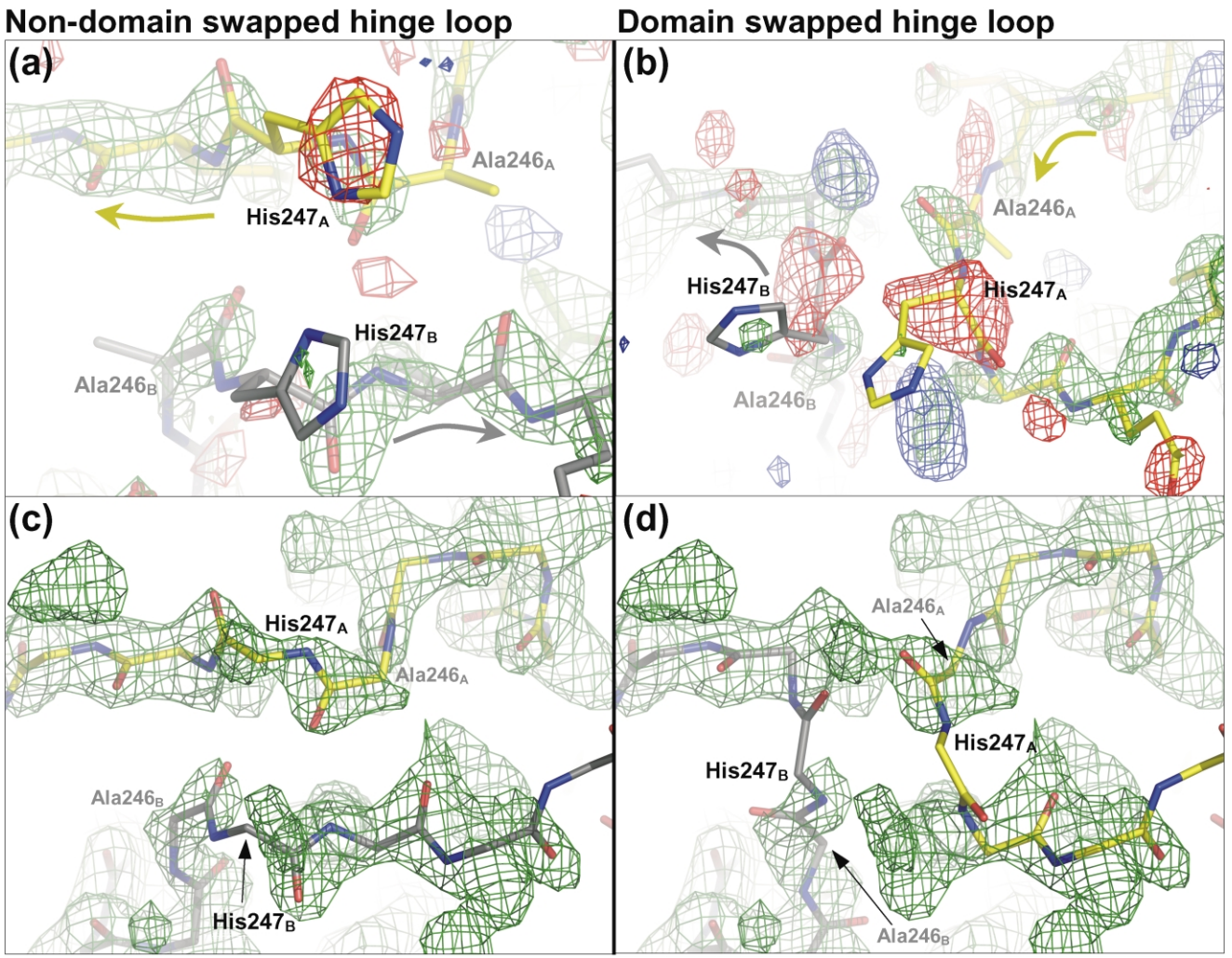

Figure 4. Crystalline *Pae* SurEα is predominantly not domain swapped. Electron density is shown for the putative hinge loop region of two refined SurEα models: (a) and (c) assuming that the C-terminal α-helix is non-domain swapped (non-DS), or (b) and (d) assuming that it is domain swapped (DS). $F_o - F_c$ maps are colored red (-3.2σ) and blue ($+3.2\sigma$), and $2F_o - F_c$ density is colored green ($+1.4\sigma$). The path of the backbone (from N to C terminus) is indicated by arrows in (a) and (b), and side-chains are omitted from (c) and (d) for clarity. Monomers A and B are distinguished by yellow (A) or gray (B) coloring of carbon atoms, and by subscript letters for the labeled hinge loop residues Ala246 and His247. The positive (blue) and negative (red) $F_o - F_c$ densities for the hinge loop show that neither the non-DS (a) nor DS (b) model is perfectly accurate, although the non-DS model is better: there is less negative $F_o - F_c$ density for the mainchain atoms of Ala246 and His247 in (a) compared to (b). The $F_o - F_c$ simulated annealing omit maps in (c) and (d) ($+3.0\sigma$) also show that the model of a non-DS dimer (c) fits this unbiased density better than the DS model (d). These electron densities, along with details discussed in the text, suggest that crystalline *Pae* SurEα is an inhomogeneous mixture of DS and non-DS states, with the non-DS form predominating.

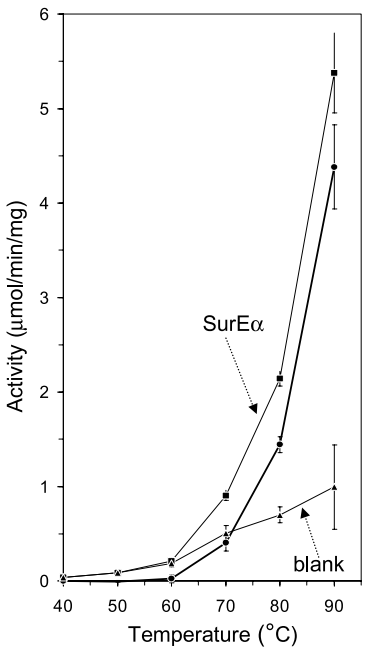

Figure 5. Temperature-dependence of *Pae* SurEα phosphatase activity on *p*-nitrophenyl phosphate (PNPP). *Pae* SurEα demonstrates minimal activity at the lowest temperature assayed (40 °C), with increasing activity up to at least 90 °C. Assays were done as described in Materials and Methods, with the enzyme incubated with 15 mM $MgSO_4$ and 15 mM disodium PNPP substrate at the indicated temperature. Each datum point is the average of three trials with *Pae* SurEα (2.6 μg, squares) or a buffer-only control (triangles). The thicker line (circles) represents the difference between these values, i.e. the SurEα specific activity (error bars indicate ±1 standard deviation for error sizes larger than the symbol).

Table 2. Comparison of *Pae* and *Tma* SurE phosphatase activities on a variety of substrates

	Substrate specificity				
	P. aerophilum		*T. maritima*		Relative activity[a] (*Pae*/*Tma*) × 100
Substrate	Activity[b]	%α-NP[c]	Activity	%α-NP	
α-Naphthyl phosphate	6 ± 1	100	127 ± 33	100	4
p-Nitrophenyl phosphate	0.5 ± 0	8	15 ± 0	12	3
5′-AMP	56 ± 16	982	63 ± 11	49	89
5′-GMP	45 ± 11	800	94 ± 17	74	49
5′-CMP	8 ± 2	135	22 ± 12	17	35
5′-TMP	13 ± 2	222	31 ± 10	25	40
2′-Deoxy-5′-GMP	30 ± 6	525	111 ± 16	87	27
2′-Deoxy-5′-AMP	47 ± 8	828	61 ± 9	48	77

[a] The relative activity column shows, for each substrate, the percentile activity of *Pae* SurEα *versus Tma* SurE.
[b] All activities are shown in μmol min^{-1} mg^{-1} ±1 standard deviation for reactions at 80 °C, using the given substrate at a final concentration of 15 mM.
[c] The %α-NP column lists the activity of the *Pae* or *Tma* enzyme as a percentage of the activity on α-naphthyl phosphate substrate.

of atoms in these residues were refined as groups in the program CNS. Throughout the course of crystallographic refinement, calculation of $2F_o - F_c$, $F_o - F_c$, and simulated annealing omit maps from models with these refined occupancies, together with the values of these occupancies, made it apparent that the crystal does not consist entirely of the non-DS (Figure 4(a) and (c)) or the DS (Figure 4(b) and (d)) conformer. Thus, evidence for an inhomogeneous crystalline mixture of DS and non-DS states is twofold: (i) neither conformation alone could be fit satisfactorily into the various maps mentioned above; and (ii) refinement of occupancies for hinge loop residues invariably led to values significantly less than 1 (but greater than 0.5). Attempts to refine alternate conformations of the hinge loop in a hybrid DS/non-DS model provided little improvement over the non-DS model, suggesting that the majority of the crystal contains non-DS dimers. Also, the non-DS model agrees more closely with relatively unbiased $F_o - F_c$ simulated annealing omit maps (compare Figure 4(c) and (d)). Parallel refinement of DS and non-DS models of *Pae* SurEα reinforced the conclusion that SurEα is (mostly) non-DS (compare Figure 4(a) and (b)). Final refinement efforts led to a model consisting exclusively of the non-DS conformer, with occupancies for the hinge loop residues (and all other atoms except for selenium) set to 1. We note that stereochemically reasonable models of SurEα with no $\phi-\psi$ dihedral violations can be built for both the DS and non-DS conformations of the hinge loop region (data not shown). Because of the ambiguity in building the hinge loop residues, and to substantiate a non-DS model for the SurEα dimer (*versus* the DS *Tma* SurE dimer), the final non-DS *Pae* structure underwent extensive model validation with the programs ERRAT,[12] PROCHECK,[13] and Verify3D[14] (Figure 1 of the Supplementary Material).

Characterization of *Pae* SurEα acid phosphatase activity

The acid phosphatase activity of *Pae* SurEα was assayed under several conditions, including various temperatures (Figure 5), with various divalent metal ion cofactors (Figure 6), and with different substrates (Table 2). At acidic conditions (pH 5.7, 80 °C) with the generic substrate *p*-nitrophenyl phosphate (PNPP), *Pae* SurEα exhibits a strongly temperature-dependent phosphatase activity that is maximal at ≥90 °C (Figure 5). The specific activity of *Pae* SurEα with Mg^{2+} cofactor at 90 °C is measured as mol phosphate released per milligram of SurE per minute, and is approximately 4.5. However, *Pae* SurEα phosphatase activity depends critically on the identity of both the divalent metal cofactor and the substrate, and Mg^{2+} is apparently not the best cofactor for the *Pae* enzyme: optimal activity on PNPP substrate at 80 °C is found with cobalt, and decreases in the

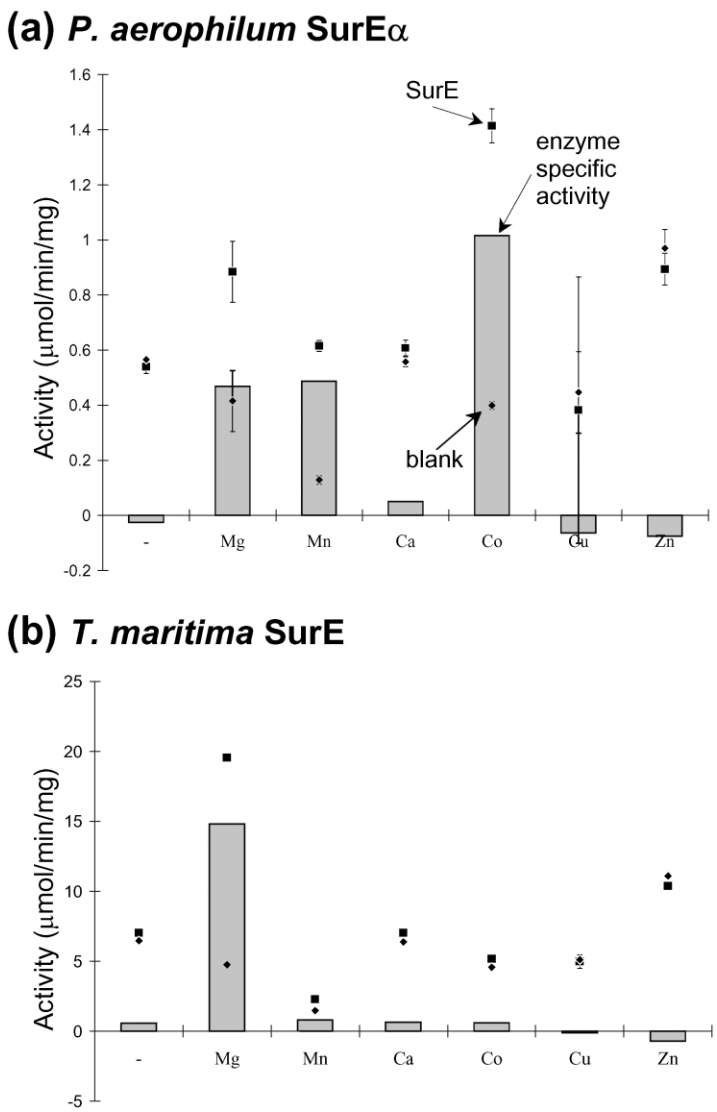

Figure 6. *Pae* SurEα phosphatase activity depends on metal ion cofactor. *Pae* SurEα activity requires a divalent metal ion cofactor, but is less specific than *Tma* SurE. Assays were performed as described in Materials and Methods. Either 2.6 μg of *Pae* SurEα (a, squares), 0.14 μg of *Tma* SurE (b, squares), or a buffer-only control (diamonds) was incubated with the indicated metal ion (15 mM) and 15 mM disodium PNPP substrate at 80 °C (each datum point is the average of three trials; error bars indicate ±1 standard deviation for error sizes larger than the symbol). Gray-shaded columns represent the enzyme-specific activity, i.e. $\text{activity}_{SurE} - \text{activity}_{\text{blank}}$.

order $Co^{2+} > Mg^{2+} \approx Mn^{2+} > Ca^{2+} > Cu^{2+} \approx Zn^{2+}$ (Figure 6(a)). In contrast, we found that *Tma* SurE is optimally active on PNPP substrate with Mg^{2+}, and is essentially inactive with these other divalent metals (Figure 6(b)). The 10–13% activity of *Tma* previously measured with Ca^{2+} may not be detected under our conditions.[7,8] A comparison of *Pae* and *Tma* SurE substrate specificities at 80 °C is shown in Table 2; these substrate specificity assays utilized Mg^{2+} for both the *Pae* and *Tma* enzymes, as Mg^{2+} is the preferred divalent metal cofactor for *Pae* SurEα with higher activity substrates such as the purine nucleoside monophosphates (NMPs). Within experimental error, *Pae* SurEα has the greatest phosphatase activity on guanosine 5′-monophosphate (GMP), adenosine-5′-monophosphate (AMP), and 2′-deoxy-5′-AMP (dAMP). But the substrate specificity of *Tma* SurE is not as clear: like the *Pae* enzyme, *Tma* is more active on the four (d)A/GMP purine (deoxy)nucleotides than the pyrimidine ones; however, it is significantly more active on the generic substrate α-naphthyl phosphate (α-NP) than any of these NMPs, and is over 20 times as active as *Pae* SurEα with α-NP (Table 2).

The putative SurE active site and a GMP-bound model

The putative active site for the acid phosphatase activities of *Pae* and *Tma* SurE is highly acidic and strongly conserved. Present as a surface pocket on the *Pae* (Figure 7(a)) and *Tma* (Figure 7(b)) structures, this site is formed primarily by conserved residues that lie near the C-terminal loops of the β-strands that form the Rossmann-fold core of SurE (Figures 1 and 2). The N-terminal Asp_8Asp_9 motif forms the center of the active site, and this same sequence and 3D structural motif is found in *Tma* SurE as in *Pae* SurEα. Mapping of residue conservation scores derived from multiple sequence alignments over the entire SurE family onto the *Pae* SurEα surface reveals that this putative active site is probably conserved in all SurEs (Figure 7(c)). GMP was chosen as a ligand because it is one of the best overall substrates for *Pae* and *Tma*

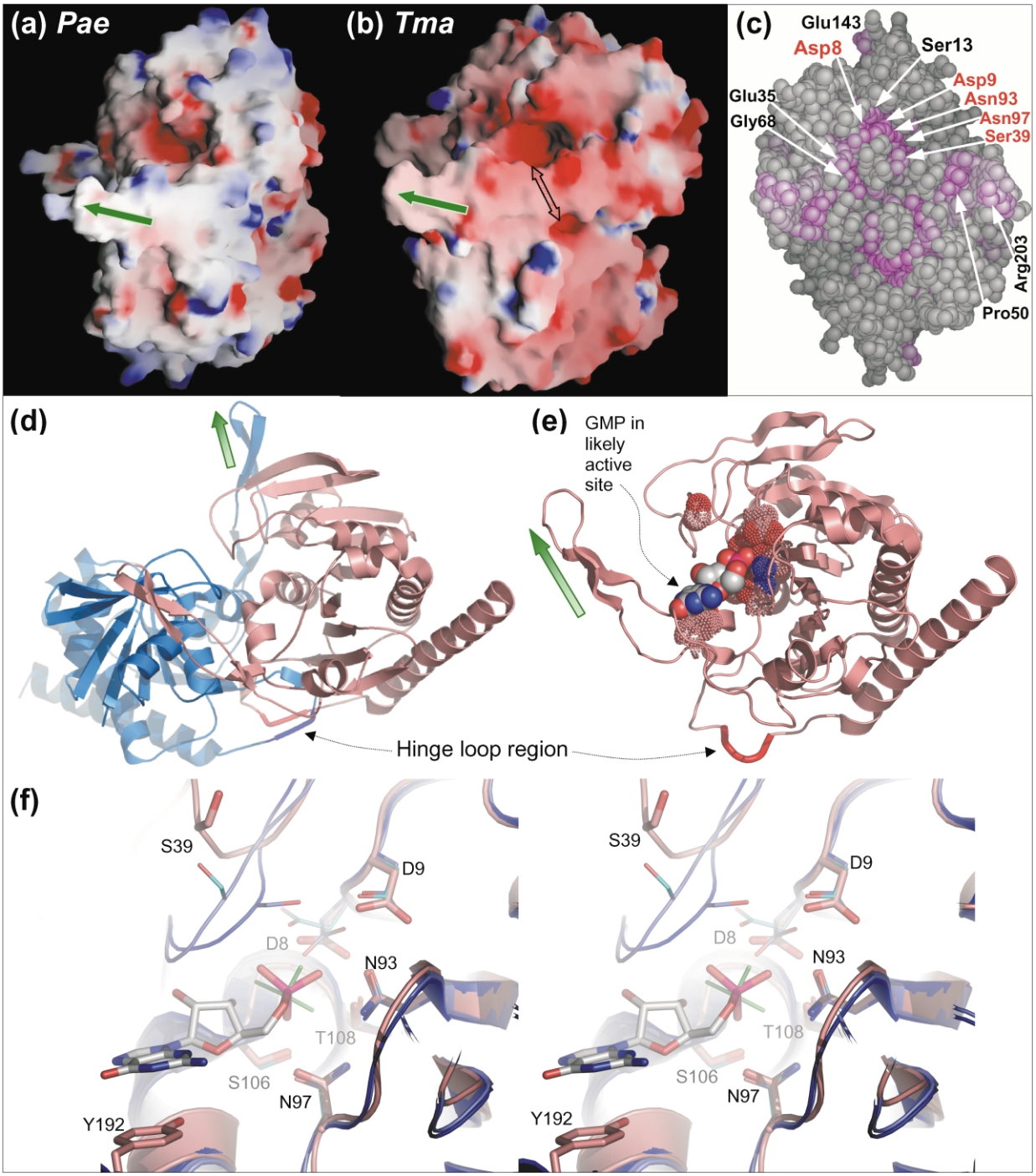

Figure 7. The putative SurE active site is highly acidic, strongly conserved, and provides a model for GMP-binding. The molecular surfaces of (a) *Pae* SurEα and (b) *Tma* SurE are displayed, colored by the calculated electrostatic potential ($-10.7kT$ (red) to $+8.6kT$ (blue) for *Pae*, and $-12.6kT$ to $+9.7kT$ for *Tma*). The dimers are oriented similarly, with the C-terminal β-hairpins (indicated by green arrows in each panel) pointing towards the left. The putative SurE active site is the highly acidic, concave surface seen in the upper left of both structures. A space-filling model of the *Pae* SurEα dimer is shown in (c), viewed down the 2-fold NCS axis (rotated roughly 90° with respect to (a) and (b)). Conserved residues are colored magenta, with the intensity of coloring reflecting the degree of conservation (likely active-site residues are labeled in red). The substantial structural conservation of the putative *Tma* and *Pae* SurE active sites is shown at increasing magnifications from (d) to (f). The *Pae* SurEα dimer is illustrated in ribbon format in (d) and the likely active site of one subunit is shown more closely in (e), along with GMP (in space-filling) and active-site residues (as surface dots). The *Pae* and *Tma* SurE active sites are superimposed in the stereoview of (f). Side-chains for the two *Tma* structures[7,8] are shown in blue and cyan, and *Pae* SurEα is colored pink (thicker sticks). Except for the Ser39 loop, the active-site structures are nearly identical. Docking of a known *Tma* SurE substrate (GMP) results in a reasonable model in which the phosphate moiety is bound in a manner identical with that of the inhibitory vanadate (shown in green) found by Lee *et al*.[7] The guanine ring stacks above the highly conserved Tyr192 side-chain. Additional protein–GMP contacts and solvent molecules are not shown, for clarity.

SurEs (Table 2), and it was docked into the active site in order to create a substrate-bound model (Figure 7(d)–(f)) in which the phosphate moiety coincides with the binding site of the inhibitory tungstate described by Lee *et al.*[7] The most striking attribute of the electrostatic surface potential of both *Tma* and *Pae* SurEα is also the most conserved feature: there is a relatively large acidic cleft that forms the putative SurE active site and continues as a narrow anionic channel in *Tma* SurE (Figure 7(a) and (b)). An interesting feature that is not shared between the electrostatic surface potentials of *Pae* and *Tma* is the significant amount of basal level anionic charge over most of the *Tma* SurE surface (Figure 7(b)); note that most of the protein surface would be neutralized only at the acidic pH values at which *Tma* SurE is optimally active.

Phylogeny and genomic organization of *surE* genes

In order to understand the phylogenetic distribution and possible evolution of SurEs, we compiled a comprehensive list of all open reading frames (ORFs) with sequence similarity to *Pae* SurEα. The database of 43 putative SurEs consists of 32 eubacterial sequences, four eukaryotic sequences, and seven archaeal proteins, represented by a total of 39 species (including the extremely radioresistant eubacterium *Deinococcus radiodurans*). We found examples of organisms with more than one *surE* gene: the genomes of *Synechocystis* sp. PCC 6803, *Nostoc* sp. PCC 7120, *A. thaliana*, and *P. aerophilum* each contain pairs of SurE paralogs, which we designate SurEα and SurEβ (Figure 8 and Table 2 of the Supplementary Material). Also, larger proteins containing SurE-like modules were found, such as a >700-residue *S. cerevisiae* ORF that contains an N-terminal half homologous to SurE and a C-terminal domain that is homologous to tubulin-tyrosine ligase (which may be phosphorylated in its Mg^{2+}/ATP-binding domain[15]).

Phylogenetic analysis of the SurE family illuminates the inter-genomic distribution of *surE* genes, but what about features of the intragenomic organization of *surE* genes, especially in terms of the possibility that they cluster with archaeal homologs of other stress-survival genes? Also, what are the gene neighbors of *Pae* SurEα and SurEβ, and where do any *pcm*, *rpoS*, or *nlpD*-like genes lie in the archaeal genomes (e.g. is there any operon-like clustering)? These questions were addressed by: (i) examination of the ORFs encoded in all six reading frames upstream and downstream of all the archaeal *surE* genes (±2500 bp); and (ii) using sequence similarity searches to locate archaeal homologs of the *pcm*, *rpoS*, and *nlpD* genes. We found no strong sequence homolog of *rpoS* or *nlpD* genes in *Pae* or any other archaea: *nlpD* homologs were found only in the eubacteria, and *rpoS* homologs could be found only in eubacteria and a few eukaryotes (primarily of the plant lineage *Viridiplantae*).

Discussion

Significance of domain swapping in SurEs

Crystalline *Pae* SurEα apparently exists as a mixture of DS and non-DS dimers, and the general significance of domain swapping for SurEs can be assessed by considering the various dimerization states of *Pae* and *Tma* SurEs. Because there is no evidence for an independently stable, closed monomer form of *Tma* SurE *in vitro*, the exchanged C-terminal α-helix classifies the *Tma* dimer structure as a candidate for 3D domain swapping, adopting the nomenclature of Schlunegger *et al.*[16] The *Pae* SurEα structure reported here provides evidence for both non-DS dimers (composed of "closed" monomers) and DS dimers (composed of "open" monomers) in the same crystal. Although crystallographic evidence for the non-DS conformer is stronger, this mixture of both states makes the SurE proteins a *bona fide* example of domain swapping. Apparently, domain swapping is a feature of SurE proteins, but its specific function is not known.

Existence of DS and non-DS states in a single crystal has implications for the significance and energetics of domain swapping. Another example of a crystalline mixture of domain swapped states was reported recently by Zhang and co-workers for the 64 residue B1 domain of protein L.[17] However, in that case the non-DS protein is naturally monomeric, and the homogeneity of the crystalline mixture allowed DS and non-DS dimers to be distinguished among the four monomers in the asymmetric unit. Nonetheless, the B1 domain and SurEα examples support the hypothesis that protein oligomers may evolve from monomers by passing through a DS stage,[18] and the possibility of mixed DS/non-DS states in a single crystal seems plausible for other DS oligomers. The DS version of a given oligomer is likely to be more thermodynamically stable than the non-DS version of the same oligomer, because of the much increased interaction surface in the DS *versus* non-DS form (see Figure 2 of the Supplementary Material for the *Pae* SurEα example). Therefore, observation of mixtures of DS/non-DS states in crystals (which take days to months to form) implies a large kinetic barrier to swapping; such a barrier may correspond to opening of the closed interface to yield an open monomer. The *Pae* SurEα structure is an extreme example of this, with the bulk of the crystal containing non-DS dimers.

Similarities and differences between *Pae* and *Tma* SurE monomers

Comparative structural analyses of *Pae* and *Tma* SurEs *via* error-scaled difference distance matrices

suggest that the conformationally variable regions comprise much of the irregular C-terminal region, including the hinge loop that connects the swapped helix to the N-terminal core. The most conformationally invariant region of *Tma* SurE is the strongly conserved N-terminal ≈170-residue core (Figures 2 and 3), and even within the ensemble of *Tma-Tma* alignments there are significant deviations in the β-hairpin and the hinge loop region that precedes the domain swapped α-helix. Whether such conformational variance is primarily due to high-amplitude fluctuations in certain dynamic regions of *Tma* SurE or discrete, slowly exchanging conformations is unclear. We found similar conformationally invariant regions in the *Tma-Pae* comparison (Figure 3(b)) and in the ensemble analysis of *Tma-Tma* (data not shown). Such consistency between *Pae-Pae* and *Pae-Tma* comparisons lends support to the interpretation of these results and to the generalization of these results to include other SurEs.

Significant differences between *Pae* and *Tma* SurE dimers and tetramers

Significant differences between *Pae* and *Tma* SurE dimers and tetramers result from several small-scale differences between *Pae* and *Tma* monomers (Figure 3) and one large-scale difference (i.e. domain swapping). There is a difference of ≈1000 Å^2 between the total buried surface area in the DS *Tma* SurE (open) dimer interface (6890 ± 40 Å^2) and the non-DS *Pae* SurEα (closed) dimer interface (5840 Å^2). The non-DS *Pae* SurEα has a much less extensive dimer interface primarily because of the non-swapped C-terminal helix (Figure 2 of the Supplementary Material). Aside from the swapped α-helix, the *Pae* and *Tma* dimer interfaces are structurally similar; there are only slight rigid-body rotations of monomers with respect to one another in the *Pae versus* the *Tma* dimer, as indicated by a 2.1 Å RMSD for alignment of *Pae* to *Tma* dimers *versus* 1.1 Å RMSD for alignment of monomers. Complete buried surface area statistics for *Pae* and *Tma* SurE monomers, dimers, and tetramers are provided in Table 1 of the Supplementary Material.

The biologically relevant oligomer of *Pae* and *Tma* SurEs is likely to be a dimer, although additional evidence suggests a tetramer. On the basis of size-exclusion chromatography and crystal packing (in which a tetramer is created by a crystallographic 2-fold), Zhang *et al.* suggested that a *Tma* SurE tetramer exists.[8] Our *in vitro* data for *Pae* SurEα reveal only a dimer, although we observe a crystalline SurEα tetramer that has the same overall structure and 222-point group symmetry as the *Tma* tetramers of Lee *et al.* and Zhang *et al*. In fact, much more surface area is buried in the dimer–dimer interface of the crystallographic *Pae* tetramer (3720 Å^2) than in either *Tma* tetramer (2220(±210) Å^2), because a slightly different conformation of the extended β-hairpins in *Pae* SurEα (Figure 3(a)) allows a much closer approach and more extensive contacts between the two dimers than is the case with the *Tma* dimers. Although *Pae* and *Tma* SurE form similar tetramers in three independent crystallization conditions (i.e. that of Lee *et al.*,[7] Zhang *et al.*,[8] and this work), the conserved active sites of one dimer are largely occluded by the C-terminal β-hairpins of the other dimer in the *Pae* and *Tma* SurE tetramers (data not shown). Thus, the functional significance of a putative SurE tetramer is unclear, and connections between the biochemical activities of *Pae* and *Tma* SurEs and the large discrepancies in their dimerization and tetramerization behavior will be of special interest.

SurEs as acid phosphatases, their conserved active sites, and a substrate-binding model

The acid phosphatase activities of *Pae* and *Tma* SurEs differ in their temperature-dependence, divalent metal cofactor requirement, and substrate specificity (Figures 5 and 6; and Table 2). Lee *et al.*[7] and Zhang *et al.*[8] found that *Tma* SurE is a novel acid phosphatase ($pH_{opt} \approx 5.5–6.2$) that is activated by Mg^{2+}, with maximal activity at ≈80 °C. *Pae* SurEα differs from the *Tma* enzyme in each of these parameters: (i) its optimal activity temperature is ≥90 °C at acidic pHs (pH 5.7 at 80 °C); (ii) whereas Mg^{2+} is the preferred ion for *Tma* activity, *Pae* SurEα is most active towards PNPP with Co^{2+} and may utilize Mg^{2+} or Mn^{2+}, though with less activity (Figure 6); and (iii) whereas the best substrate found for *Tma* SurE is the generic phosphate ester α-naphthyl phosphate (α-NP), *Pae* SurEα is more active on purine (deoxy)nucleoside monophosphates, particularly the substrates GMP, AMP, and dAMP. The difference in optimal activity temperatures is consistent with the fact that *Tma* is a thermophilic eubacterium that has an optimal growth temperature of 80 °C, whereas the hyperthermophilic *Pae* thrives at 100 °C. Elucidation of the differences between *Pae* and *Tma* metal cofactor requirements and substrate specificities is difficult because these two parameters are interconnected: for example, the relative activity of *Pae* SurEα with Mg^{2+} *versus* Co^{2+} is over five times higher with dGMP than with PNPP substrate (data not shown). In order to understand the structural basis of these phosphatase activities, we determined the likely *Pae* SurEα active site and created a substrate-bound model.

The putative SurE active site is a highly conserved cavity on the *Tma* and *Pae* surfaces (Figure 7). The likely *Tma* SurE active site was identified on the basis of mutation studies and the crystallographically located divalent metal ion-binding sites.[7,8] The most strictly conserved residues in the SurE family cluster about this active-site region (Figure 7(c)). In order to gain further insight into binding of potential substrates and catalysis by SurEs, we created a model of guanosine-5′-monophosphate (GMP) bound to the conserved *Pae*

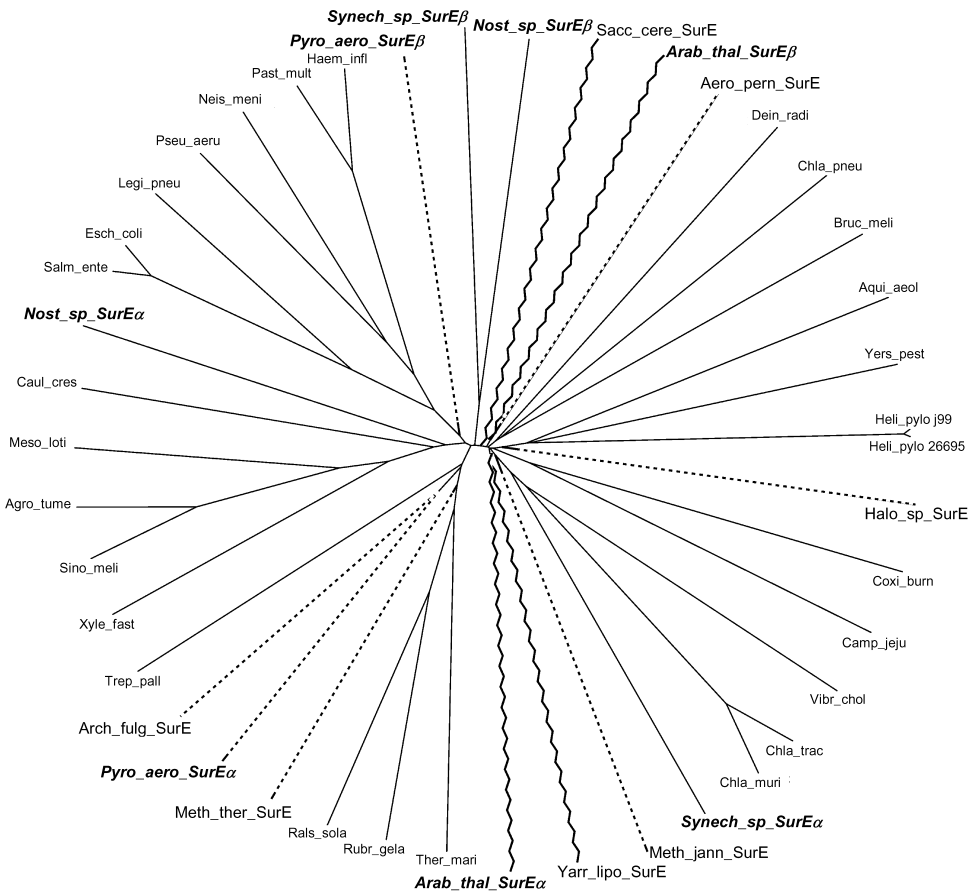

Figure 8. Phylogenetic distribution and paralogs of *surE* genes. An unrooted phylogenetic tree of SurEs is shown, as calculated from a multiple sequence alignment of all 43 detectable SurE homologs. Eukaryotic SurEs are shown as pleated lines and archaeal species as broken lines; the remaining SurEs are eubacterial. Paralogous SurE pairs (α, β) are italicized. Note the great dispersion in the SurE family amongst the eubacteria, eukaryotes, and archaea, and that paralogous *surE* genes do not cluster into clades.

SurEα active site (Figure 7). In addition to showing that the *Pae* SurEα active site can accommodate GMP, our model of a SurE·GMP complex elucidates the importance of many conserved active-site residues. The *Tma* SurE protein binds a divalent metal ion (Mg^{2+} or Ca^{2+}) in the active site such that the metal is chelated by several conserved residues, including the strictly conserved Asp8Asp9 pair. A water molecule is tightly bound 2.2 Å away from Mg^{2+} in the structure described by Lee *et al.*,[7] and this polarized water molecule may serve as a nucleophile for attack on the substrate's phosphate center. As *Pae* SurEα could be crystallized only in the presence of EDTA, no divalent metal ions were found at the active site. However, water molecules were found in the two sites just described: one water molecule occupies the same divalent metal ion-binding site as in *Tma*, and the other is hydrogen bonded to this one. Both of these water molecules are compatible with the location of the modeled GMP. Additionally, specific SurE···GMP contacts reveal likely features of substrate recognition, e.g. a possible π-stacking interaction between the highly conserved Tyr192 side-chain and the guanine ring of GMP (Figure 7(f)). This interaction could explain the specificity of *Pae* SurEα for purine NMPs, since a pyrimidine ring would not extend far enough from the Asp8Asp9-metal center to stack upon the phenyl side-chain of Tyr192. The primary endogenous substrate for *Pae* SurEα and other SurEs is unknown. In addition to a possible role in phosphorous scavenging during stressful conditions,[8] our results suggest that SurEs may have a nucleic acid substrate, such as the 5′ phosphate group of some RNA species.

Phylogenetic distribution of *surE* genes

An unrooted phylogenetic tree was inferred by the application of distance matrix methods to multiple sequence alignments of the 43 known SurE sequences, and shows a large dispersion in the SurE lineages. That is, the tree displays very few bifurcated nodes, and most of the SurE sequences cannot be grouped into clades. The tree shows that the SurE sequences do not cluster by kingdom: archaeal SurEs are interspersed with eukaryotic and eubacterial ones in an apparently random way (Figure 8). The

phylogenetic relationships of SurE paralog pairs, such as *Pae* SurEα/β or *Nostoc* SurEα/β suggest that the second member of these SurE pairs may not have arisen by gene duplication and neutral drift within these genomes. If gene duplication led to two SurEs in these genomes, the two paralogs would likely have a greater degree of sequence similarity and would share a stronger phylogenetic similarity than that shown in Figure 8. For example, the two SurEs from two subspecies of *Helicobacter pylori* are closely related, as are the *E. coli* and *Salmonella enterica* SurEs; however, members of the four α/β pairs are apparently only distantly related. Thus, duplicate *surE* genes may have arisen by horizontal gene transfer. Our finding of two *surE* genes in the hyperthermophile *Pae* (which grows up to 104 °C) is especially interesting, given a recent report that the single *surE* gene in *E. coli* is duplicated in strains that are evolved for 2000 generations at elevated temperatures (the authors speculated that such duplication, along with the *pcm*, *rpoS*, and *nlpD* genes, may facilitate thermal adaptation in *E. coli*[5]).

Genomic organization in archaea and *surE* gene neighbors

The genomic organization of, or even the presence of, putative stress-survival genes is not conserved in the eubacteria and archaea. Homologs of the *rpoS* or *nlpD* genes were not found, but one significant *pcm* homolog was detected in *Pae*. However, unlike the case in several eubacterial genomes, this *pcm* is not located near either *surE* gene in *Pae*, but is nearly 0.5 Mbp away (Figure 3(a) of the Supplementary Material). The same result was found in other archaeal genomes: in each case, the nearest gene neighbors of *surE* were not homologous to the *pcm*, *rpoS*, or *nlpD* genes. Several of the ORFs adjacent to archaeal *surE* genes have no strong sequence matches to proteins of known function, being annotated as conserved hypothetical proteins. However, in some cases, a homolog of known function can be found for *surE* gene neighbors, and several of the reactions catalyzed by these homologs involve some form of phosphate ester hydrolysis. For example, the nearest gene neighbor of *Pae surE*α encodes a putative purine NTPase (≈1500 nt upstream and in the same reading frame). This is an especially salient result, given the substrate specificity of *Pae surE*α described above. A homolog of CTP-synthase is ≈1200 nt upstream of, and in the same reading frame as, *Pae surE*β (Figure 3 of the Supplementary Material). Other examples of archaeal gene neighbors are: (i) a putative protein tyrosine phosphatase (PTP) upstream of *Methanobacterium thermautotrophicum surE* (Figure 3(c) of the Supplementary Material); (ii) a homolog of ribose-5′-phosphate isomerase downstream of (and overlapping) the *Archaeoglobus fulgidus surE* gene; and (iii) an adjacent dihydropteroate synthase (DHPS) gene in *Aeropyrum pernix* (encoded in a reverse reading frame). In a reaction analogous to that catalyzed by DHPS, the nearest *Pae* SurEβ gene neighbor (CTP-synthase) mediates the condensation of UTP and an amino group to form cytosine 5′-triphosphate. Therefore, in many cases the predicted biochemical activities of *surE* gene neighbors can be reconciled with the acid phosphatase activities of *Pae* and *Tma* SurEs, and may help to place these activities in broader metabolic contexts.

Conclusions

Until the recent reports by Lee *et al.*[7] and Zhang *et al.*[8] for *Tma* SurE, there was no structural or biochemical knowledge about the SurE family and its *in vivo* function. In order to extend and generalize their results, we determined the crystal structure of *Pae* SurEα to 2.0 Å resolution. The *Pae* and *Tma* monomers adopt similar structures, consisting of N-terminal Rossmann-like folds and irregular, C-terminal domains that mediate oligomerization. Crystalline *Pae* SurEα differs from *Tma* SurE in that it apparently forms a mixture of domain-swapped and non-domain-swapped dimers, with the non-domain-swapped form predominating. This shows that SurE proteins can exist in both monomeric and dimeric forms, and suggests that the transition could be of functional significance. More minor differences in the two structures were revealed by the application of error-scaled difference distance matrices. The considerable structural similarity of the SurE active sites allowed us to model the *Pae* enzyme bound to a potential substrate (GMP), and the SurE·GMP model suggests the importance of conserved active site residues. Characterization of the temperature-dependence, substrate specificity, and divalent metal ion requirements of the *Pae* SurEα acid phosphatase activity suggests that the *Pae* and *Tma* enzymes probably have similar (but not identical) functions. Analyses of the phylogeny and genomic organization of SurE reveal examples of genomes with multiple *surE* genes, and suggest a generic phosphatase-like function for other members of the SurE family.

Materials and Methods

Cloning, expression, and purification of *Pae* SurEα

A genomic phosmid clone containing the *Pae* SurEα open reading frame was kindly provided by the laboratory of Jeffrey H. Miller (UCLA). Using its DNA sequence, we designed primers for polymerase chain reaction (PCR) amplification. Blunt-end PCR products were cloned into a pET-22b(+) expression vector (Novagen) *via* intermediate subcloning into the pCR-Blunt vector (Invitrogen). Ligations were transformed into chemically-competent *E. coli*, and DNA sequencing verified that recombinant protein would be wild-type (wt) SurEα with the following C-terminal 14-residue His_6-tag: SKLAAALEHHHHHH. Due to the relative rarity of the AGG and AGA arginine codons in *E. coli*, *Pae*

SurEα over-expression required co-transformation with a $tRNA^{Arg}$-encoding vector (see the rare codon calculator†). Recombinant SurEα was over-expressed in BL21(DE3) *E. coli* at 37 °C by standard protocols using 1 mM isopropyl-β-D-thiogalactoside (IPTG) induction of the T7*lac*-based promoter. Approximately 10 mg of soluble protein was expressed per liter of cell culture. Selenomethionine (SeMet)-substituted SurEα was prepared in exactly the same way as the native/wt protein, except that the expression was performed in M9 minimal media supplemented with SeMet (as described by Van Duyne *et al.*[19]).

Harvested cells were thawed and re-suspended in a high-salt concentration buffer (20 mM NaHepes (pH 7.8), 1.5 M NaCl, 0.5% (v/v) Triton X100, 30 mM PMSF), and cells were lyzed by a combination of lysozyme treatment and French-press. Initial protein purification was achieved by heating the cleared supernatant to ≈80 °C (>85% purity as estimated by density scans of SDS-PAGE lanes), followed by high-speed centrifugation to remove the bulk of denatured *E. coli* proteins. Recombinant SurEα-His6x was further purified on a Ni^{2+}-charged iminodiacetic acid-Sepharose column. *Pae* SurEα was eluted in an imidazole gradient, and it may be significant that yellow-colored fractions from an earlier point in the gradient reproducibly contained a single protein of ≈20 kDa. Affinity chromatography resulted in >99% pure protein as estimated by several independent techniques (SDS-PAGE, matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry, electrospray mass spectrometry, and gel-filtration chromatography). These methods verified that the final purified protein consists of full-length wt SurEα with the appended 14 residue His-tag. Mass spectrometry was used to verify SeMet incorporation. After chromatography, all attempts to exchange SurEα-His_6 into a buffer incapable of chelating divalent metal ions (e.g. any buffer lacking imidazole or EDTA) were unsuccessful; the protein would invariably precipitate out of solution, presumably due to His-tag-mediated polymerization in the presence of divalent cations. Therefore, for crystallization efforts, SeMet-containing *Pae* SurEα was dialyzed into 10 mM Tris–HCl (pH 8.0), 5 mM EDTA.

Crystallization and X-ray data collection

SeMet-labeled SurEα was concentrated to ≈11 mg/ml, and hanging-drop vapor-diffusion and sparse matrix screening yielded several initial crystallization leads with different habits. These leads were extensively optimized by varying crystallization parameters, particularly protein and precipitant (polyethylene glycol, PEG) concentrations. In terms of a crystallization response surface,[20] the most critical parameter was the sampling of a different region of crystallization space; in this case, *via* the addition of a reductant (dithiothreitol, DTT). The efficacy of adding 10 mM DTT to the crystallization condition is rationalized easily, because biophysical characterization of *Pae* SurEα shows that it forms disulfide-bonded dimers *in vitro* (unpublished results). The final, optimized crystallization conditions for trigonal SurEα crystals grown in hanging-drops at 19.8 °C are: 5 μl drops (2.5 μl well + 2.5 μl of 11.4 mg/ml of SeMet SurEα) over 600 μl wells (0.083 M Tris (pH 8.55), 21.7% (v/v) PEG-4000, 0.17 M sodium acetate, 15% (v/v) glycerol). A single SeMet crystal of reasonable diffraction quality appeared within one year, and grew as trigonal prisms of maximum size ≈0.2 mm × 0.4 mm (the native protein never crystallized).

Auto-indexing and scaling of diffraction patterns revealed the space group to be either $P3_121$ or $P3_221$, with unit cell dimensions $a = 90.5$ Å, $c = 129.95$ Å. These cell dimensions and the molecular mass of *Pae* SurEα-His6x (30,733.1 Da) suggested a dimer in the asymmetric unit. For $Z = 12$ monomer/cell, the calculated Matthews coefficient ($V_M = 2.50$ Å^3/Da) corresponds to a solvent content of 50.8% (v/v). Final SeMet-MAD data sets were collected on an ADSC Quantum-4 charge-coupled device (CCD) detector at ALS beamline 5.0.2. Data were collected on the crystal in a cryogenic nitrogen stream at −168 °C (105 K). All images were indexed, integrated, and reduced in DENZO, and reflections were scaled and merged in SCALEPACK.[21] Appropriate wavelengths for inflection, peak, and high-energy remote data sets were selected from X-ray fluorescence scans about the K absorption edge of the SeMet crystal (≈12.6578 keV). Three complete data sets were collected from the single SeMet crystal (Table 1).

Crystallographic MAD phasing and refinement

Two indications that the SeMet crystal would be suitable for multi-wavelength anomalous dispersion (MAD) phasing were: (i) χ^2 values >1 for the merging of I^+ and I^- reflections indicated a reasonable anomalous signal (Table 1); and (ii) large anomalous difference (ΔF^2_{ano}) Patterson peaks for the data set collected at the selenium peak wavelength. The integrated Patterson and direct methods program SHELXD‡ located eight Se sites per asymmetric unit (out of 12 expected sites), using a single-wavelength anomalous scattering approach with the peak data set. Sites were verified by comparing predicted self-vectors and cross-vectors with observed peaks in anomalous difference Patterson maps. The program MLPHARE[22] was used for maximum likelihood heavy atom and MAD phase refinement to 2.8 Å (that being the high-resolution limit of the high-energy remote data set). Phases for the centrosymmetric solution were also calculated and refined; i.e. the inverted hand of the Se positions in the enantiomorphic space group ($P3_221$). Density modification with the program DM[23] distinguished the correct enantiomorph of Se sites and improved electron density map quality. Maps calculated from experimental phases were of excellent quality (Figure 1 of the Supplementary Material), with protein secondary structure elements clearly identifiable. Phases were extended from 2.8 Å to 2.0 Å with DM (including 2-fold NCS averaging). Rigid secondary structure elements were initially fit into 2.5 Å resolution maps automatically with the program MAID,[24] and this served as a useful starting point for automated model building of ≈87% of the protein backbone (485 out of 560 residues/dimer in 13 chains) with the program ARP/wARP.[25]

Manual model building was done in O,[26] and the program CNS[27] was used for model refinement. Refinement proceeded by standard protocols, using the maximum-likelihood target function for amplitudes (mlf), bulk solvent correction, and anisotropic scaling correction terms. Initially, the two monomers in the asymmetric unit were refined with only weak NCS restraints imposed, and for final rounds of refinement the two

† http://www.doe-mbi.ucla.edu/cgi/cam/racc.html

‡ http://shelx.uni-ac.gwdg.de/SHELX/

monomers were refined independently. Refinement of individual atomic positions, isotropic temperature factors, and simulated annealing torsion angle dynamics was performed in most rounds. Each refinement round ended with inspection of the agreement between the model and σ_A-weighted $2F_o - F_c$, $F_o - F_c$, and simulated annealing omit maps (the latter only as necessary). Atomic occupancies for Se atoms and hinge loop residues were refined as necessary in CNS (see Results). The final model consists exclusively of the non-DS dimer, and contains 276/280 residues for one monomer, and 278/280 for the other. A total of 287 water molecules, seven glycerol molecules, and two acetate groups were modeled as solvent. Model validation utilized the programs ERRAT,[12] PROCHECK,[13] and Verify3D.[14] Experimental structure factors and the refined *Pae* SurEα model have been submitted to the PDB (code 1L5X).

Sequence and structure analyses

Homologs of *Pae* SurEα were found *via* iterative PSI-BLAST[28] searches of the most current non-redundant database of protein sequences at NCBI. This final list of 43 SurE homologs is shown in Table 2 of the Supplementary Material. Multiple sequence alignments over the entire list, as well as just the seven archaeal sequences, were performed with CLUSTALW.[29] Pairwise sequence similarity scores were calculated by the Smith–Waterman algorithm, as implemented in the GCG software package.[30] An unrooted phylogenetic tree for all 43 SurEs was inferred from the distance matrix methods in the Phylogeny inference program PHYLIP.[31]

Structural alignments were created with various programs. For example, active-site regions (which are similar in structure) were aligned with the Kabsch least-squares method in the program ALIGN,[32] whereas entire monomeric or dimeric *Pae* and *Tma* structures (which are more dissimilar) were optimally aligned by the combinatorial extension algorithm.[33] Calculations of the electrostatic potentials at surfaces were performed in GRASP,[34] and buried surface areas were calculated in CNS by the Lee & Richards method.[35] Comparative structural analyses of several SurE models was performed *via* DDMs[10] in the program ESCET.[11] To this end, the ESCET analysis was performed twice: (i) using an ensemble of the eight crystallographically independent *Tma* SurE models refined by Lee *et al.*[7] and Zhang *et al.*;[8] or (ii) using a single *Tma*/*Pae* pair of structures (e.g. a single *Tma* monomer from PDB code 1J9J and a single *Pae* monomer). In the latter case, the ESCET analysis was restricted to portions of the two chains that aligned in 3D (as determined by combinatorial extension). The programs GRASP† and PyMOL‡ were used for electron density Figures and other structural illustrations.

Phosphatase activity assays

The phosphatase activity of SurE on *p*-nitrophenyl phosphate (PNPP) substrate was assayed by measuring the increased absorbance at 410 nm that results from the removal of the phosphate group from PNPP. Reactions utilizing the disodium salt of PNPP (Sigma) were performed in a circulating waterbath at the indicated temperatures (Figure 5), using mildly acidic sample buffers that consisted of 100 mM Mops (pH 6.2 when measured at 22 °C, pH 5.7 measured at 80 °C), 5% glycerol, 15 mM the specified metal ion (Figure 6), 15 mM the specified substrate (PNPP or otherwise, see Table 2), and between 0.1 μg and 1 μg of *Pae* or *Tma* SurE protein (as indicated in the Figure legends) in a final reaction volume of 250 μl. Equivalent volumes of buffer alone were used for reference/blank reactions. At the three and eight minute time-points, 100 μl of the reaction mixture was mixed with 900 μl of water, the absorbance at 410 nm was measured, and the SurE activity was calculated as mol *p*-nitrophenol liberated (based on a molar extinction coefficient of 18,500 for *p*-nitrophenol). Phosphatase activities at 80 °C on other substrates (Table 2) were measured by quantifying the difference in free phosphate concentration between reactions with and without SurE. For these other substrates, phosphate release was measured using the EnzChek system from Molecular Probes (E-6646) as described.[8] All reactions were performed in triplicate with a matched triplicate of non-SurE controls.

† http://trantor.bioc.columbia.edu/grasp
‡ http://pymol.sourceforge.net/

Acknowledgements

We thank Dr Sorel Fitz-Gibbon (UCLA) for the phosmid vector containing the *Pae surE*α gene, as well as the Advanced Light Source at Berkeley National Laboratory for use of their synchrotron beamline 5.0.2. We thank Drs Duilio Cascio and Michael Sawaya (UCLA) for indispensable crystallographic advice, Dr Thomas Schneider (University of Göttingen) for assistance with the ESCET program, and Drs Daniel Anderson, James Bowie, and Yan-Shun Liu (UCLA) for helpful discussions. This work was funded by grants from the DOE, NIH (GM26020 and AG18000 to S.C.), and an NSF graduate research fellowship (C.M.). D.E. is an Investigator of the Howard Hughes Medical Institute.

References

1. Li, C., Ichikawa, J. K., Ravetto, J. J., Kuo, H. C., Fu, J. C. & Clarke, S. (1994). A new gene involved in stationary-phase survival located at 59 minutes on the *Escherichia coli* chromosome. *J. Bacteriol.* **176**, 6015–6022.
2. Li, C., Wu, P. Y. & Hsieh, M. (1997). Growth-phase-dependent transcriptional regulation of the pcm and surE genes required for stationary-phase survival of *Escherichia coli*. *Microbiology*, **143**, 3513–3520.
3. Loewen, P. C. & Hengge-Aronis, R. (1994). The role of the sigma factor sigma S (KatF) in bacterial global regulation. *Annu. Rev. Microbiol*, **48**, 53–80.
4. Visick, J. E., Ichikawa, J. K. & Clarke, S. (1998). Mutations in the *Escherichia coli* surE gene increase isoaspartyl accumulation in a strain lacking the pcm repair methyltransferase but suppress stress-survival phenotypes. *FEMS Microbiol. Letters*, **167**, 19–25.
5. Riehle, M. M., Bennett, A. F. & Long, A. D. (2001). Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **98**, 525–530.
6. Treton, B. Y., Le Dall, M. T. & Gaillardin, C. M. (1992). Complementation of *Saccharomyces cerevisiae* acid

phosphatase mutation by a genomic sequence from the yeast *Yarrowia lipolytica* identifies a new phosphatase. *Curr. Genet.* **22**, 345–355.
7. Lee, J. Y., Kwak, J. E., Moon, J., Eom, S. H., Liong, E. C., Pedelacq, J. D. *et al.* (2001). Crystal structure and functional analysis of the SurE protein identify a novel phosphatase family. *Nature Struct. Biol.* **8**, 789–794.
8. Zhang, R. G., Skarina, T., Katz, J. E., Beasley, S., Khachatryan, A., Vyas, S. *et al.* (2001). Structure of *Thermotoga maritima* stationary phase survival protein SurE: a novel acid phosphatase. *Structure (Camb)*, **9**, 1095–1106.
9. Fitz-Gibbon, S. T., Ladner, H., Kim, U. J., Stetter, K. O., Simon, M. I. & Miller, J. H. (2002). Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl Acad. Sci. USA*, **99**, 984–989.
10. Schneider, T. R. (2000). Objective comparison of protein structures: error-scaled difference distance matrices. *Acta Crystallog. sect. D*, **56**, 714–721.
11. Schneider, T. R. (2002). A genetic algorithm for the identification of conformationally invariant regions in protein molecules. *Acta Crystallog. sect. D*, **58**, 195–208.
12. Colovos, C. & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511–1519.
13. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–290.
14. Eisenberg, D., Luthy, R. & Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* **277**, 396–406.
15. Idriss, H. T. (2000). Phosphorylation of tubulin tyrosine ligase: a potential mechanism for regulation of alpha-tubulin tyrosination. *Cell Motil. Cytoskeleton*, **46**, 1–5.
16. Schlunegger, M. P., Bennett, M. J. & Eisenberg, D. (1997). Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Advan. Protein Chem.* **50**, 61–122.
17. O'Neill, J. W., Kim, D. E., Johnsen, K., Baker, D. & Zhang, K. Y. (2001). Single-site mutations induce 3D domain swapping in the B1 domain of protein L from *Peptostreptococcus magnus*. *Structure*, **9**, 1017–1027.
18. Bennett, M. J., Choe, S. & Eisenberg, D. (1994). Domain swapping: entangling alliances between proteins. *Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.
19. Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1993). Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229**, 105–124.
20. Carter, C. W., Jr (1997). Response surface methods for optimizing and improving reproducibility of crystal growth. *Methods Enzymol.* **276A**, 74–99.
21. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326.
22. Dodson, E. J., Winn, M. & Ralph, A. (1997). Collaborative computational project number 4: providing programs for protein crystallography. *Methods Enzymol.* **276**, 620–633.
23. Cowtan, K. & Main, P. (1998). Miscellaneous algorithms for density modification. *Acta Crystallog. sect. D*, **54**, 487–493.
24. Levitt, D. G. (2001). A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps. *Acta Crystallog. sect. D*, **57**, 1013–1019.
25. Perrakis, A., Morris, R. & Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463.
26. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, . (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A*, **47**, 110–119.
27. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W. *et al.* (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905–921.
28. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
29. Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**, 383–402.
30. Womble, D. D. (2000). GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol. Biol.* **132**, 3–22.
31. Kuhner, M. K. & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468.
32. Satow, Y., Cohen, G. H., Padlan, E. A. & Davies, D. R. (1986). Phosphocholine binding immunoglobulin Fab McPC603: an X-ray diffraction study at 2.7 Å. *J. Mol. Biol.* **190**, 593–604. http://www.doe-mbi.ucla.edu/People/Software/ALIGN.html..
33. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
34. Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281–296.
35. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Edited by D. Rees

(Received 16 August 2002; received in revised form 30 December 2002; accepted 30 December 2002)

Supplementary Material comprising three Figures and two Tables is available on Science Direct