bacteria

MDPI

Article

Comparative Analysis and Ancestral Sequence Reconstruction of Bacterial Sortase Family Proteins Generates Functional Ancestral Mutants with Different Sequence Specificities

Jordan D. Valgardson [1,†], Sarah A. Struyvenberg [1,‡], Zachary R. Sailer [2,3,§], Isabel M. Piper [1,‖], Justin E. Svendsen [1,2], D. Alex Johnson [1,¶], Brandon A. Vogel [1], John M. Antos [1], Michael J. Harms [2,3] and Jeanine F. Amacher [1,*]

[1] Department of Chemistry, Western Washington University, Bellingham, WA 98225, USA; jvalgard@stanford.edu (J.D.V.); sastruyvenberg@msn.com (S.A.S.); isabel_piper@berkeley.edu (I.M.P.); jsvends2@uoregon.edu (J.E.S.); dajohnso@caltech.edu (D.A.J.); vogelb2@wwu.edu (B.A.V.); antosj@wwu.edu (J.M.A.)
[2] Department of Chemistry and Biochemistry, University of Oregon, Eugene, OR 97403, USA; zachsailer@gmail.com (Z.R.S.); harms@uoregon.edu (M.J.H.)
[3] Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA
* Correspondence: amachej@wwu.edu; Tel.: +1-360-650-4397
† Current address: Department of Chemical and Systems Biology, Stanford University, Palo Alto, CA 94305, USA.
‡ Current address: Lumen Biosciences, Seattle, WA 98103, USA.
§ Current address: Apple Inc., Cupertino, CA 95014, USA.
‖ Current address: Department of Chemistry, University of California, Berkeley, CA 94720, USA.
¶ Current address: Department of Bioengineering, CalTech, Pasadena, CA 91125, USA.

Citation: Valgardson, J.D.; Struyvenberg, S.A.; Sailer, Z.R.; Piper, I.M.; Svendsen, J.E.; Johnson, D.A.; Vogel, B.A.; Antos, J.M.; Harms, M.J.; Amacher, J.F. Comparative Analysis and Ancestral Sequence Reconstruction of Bacterial Sortase Family Proteins Generates Functional Ancestral Mutants with Different Sequence Specificities. *Bacteria* **2022**, *1*, 121–135. https://doi.org/10.3390/bacteria1020011

Academic Editor: Bart C. Weimer

Received: 29 April 2022
Accepted: 7 June 2022
Published: 9 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: Gram-positive bacteria are some of the earliest known life forms, diverging from gram-negative bacteria 2 billion years ago. These organisms utilize sortase enzymes to attach proteins to their peptidoglycan cell wall, a structural feature that distinguishes the two types of bacteria. The transpeptidase activity of sortases make them an important tool in protein engineering applications, e.g., in sortase-mediated ligations or *sortagging*. However, due to relatively low catalytic efficiency, there are ongoing efforts to create better sortase variants for these uses. Here, we use bioinformatics tools, principal component analysis and ancestral sequence reconstruction, in combination with protein biochemistry, to analyze natural sequence variation in these enzymes. Principal component analysis on the sortase superfamily distinguishes previously described classes and identifies regions of relatively high sequence variation in structurally-conserved loops within each sortase family, including those near the active site. Using ancestral sequence reconstruction, we determined sequences of ancestral *Staphylococcus* and *Streptococcus* Class A sortase proteins. Enzyme assays revealed that the ancestral *Streptococcus* enzyme is relatively active and shares similar sequence variation with other Class A *Streptococcus* sortases. Taken together, we highlight how natural sequence variation can be utilized to investigate this important protein family, arguing that these and similar techniques may be used to discover or design sortases with increased catalytic efficiency and/or selectivity for sortase-mediated ligation experiments.

Keywords: sortases; enzymes; protein engineering; principal component analysis; network analysis; bioinformatics; ancestral sequence reconstruction; evolution

1. Introduction

Gram-positive bacteria accounted for 76% of all bloodstream infections in 2000, up from 62% in 1995 [1]. Although varied by region and over time, these numbers have stayed relatively consistent for the past 20 years [2–4]. These organisms are defined in part by their thick peptidoglycan layer as compared to gram-negative bacteria, which they

diverged from roughly 2 billion years ago [1,5,6]. Sortase enzymes are critical for the ability of gram-positive bacteria to attach proteins to the cell exterior, as well as to build the pili [7–10]. Due to this activity, sortases are a potential therapeutic target for antibiotic development, and they are actively-used tools for protein engineering [11,12]. Several of the infections mentioned above are caused by pathogenic *Staphylococci* and *Streptococci*, e.g., *Staphylococcus aureus* and *epidermidis*, and *Streptococococcus pneumoniae, pyogenes,* and *agalactiae* [1]. Therefore, a greater understanding of proteins from these organisms may prove valuable in the fight against gram-positive bacterial infection.

There are six main classes of sortase (class A-F); the first-characterized and best-studied bacterial sortase is the Class A sortase from *Staphylococcus aureus* (saSrtA) [13]. This enzyme recognizes the Cell Wall Sorting Signal (CWSS) sequence LPXTG, where X=any amino acid. Following cleavage of the initial protein target, an acyl-enzyme intermediate is formed. A secondary substrate then acts as a nucleophile, and a final ligation product is generated [9]. Peptidase activity occurs between the Thr and Gly residues, and positions are defined as P4 = Leu, P3 = Pro, P2 = X, P1 = Thr, and P1′ = Gly. Other Class A sortases, e.g., *Streptococcus pyogenes* SrtA (spySrtA), are predicted to contain a closely related recognition mechanism, and our group recently showed that recognition of the P1′ residue is partially mediated by residues in the β4-β5 and β7-β8 loops, highlighting the importance of these conserved structural features [14,15].

The catalytic activity of sortases make them an exciting tool in protein engineering, where sortase-mediated ligation (SML) or *sortagging* techniques are commonly employed to create a variety of products, including the recent development of an in vivo assay using engineered saSrtA to label amyloid-β protein in human cerebrospinal fluid and the implementation of ligation site switching to allow assembly of multiple fragments using a single sortase enzyme, amongst many others [11,16–18]. Despite their uses, sortagging applications are hindered by the poor relative enzymatic efficiency of saSrtA and other naturally occurring sortases studied to date [19–21]. Directed evolution studies performed in 2011 were successful in generating a saSrtA pentamutant (P94R/D160N/D165A/K190E/K196T) with an overall catalytic efficiency increase >100-fold [21]. Engineering of additional variants of saSrtA and other Class A sortases is an area of ongoing work. An example includes the incorporation of two additional mutations to the saSrtA pentamutant at the calcium-binding site, which led to a calcium-independent saSrtA heptamutant [20,22–24]. Other studies use directed evolution or other engineering techniques to alter the substrate specificity of saSrtA, e.g., a recent study that reported an saSrtA variant which recognizes an LMVGG substrate motif in the amyloid-β protein [17].

Variation in substrate selectivity also naturally exists amongst bacterial sortases. Although saSrtA is selective for the LPXTG target sequence, this is not true of all Class A sortases. Work from ourselves and others revealed that other Class A sortases can recognize a variety of amino acids at multiple positions [14,15,25,26]. A complete understanding of the selectivity determinants of these alternate preferences is not known. Furthermore, there are six known classes of sortases (A–F). Many of these classes share a similar recognition motif as Class A sortases, including Classes C-F (Class C: [I/L][P/A]XTG; Class D: LPNTA; Class E: LAXTG; Class F: less is known, but it is likely similar to SrtA, LPXTG) [27,28]. However, the recognition motif of Class B sortases is NP[Q/K]TN [27]. Taken together, we hypothesize that investigating sequence variation of individual classes of sortases, as well as the sortase superfamily, may identify sortases with improved catalytic efficiency and/or unique recognition motifs.

Ancestral sequence reconstruction (ASR) is a powerful technique that combines our growing knowledge of the proteomes of extant organisms with statistical methods in order to predict the sequences of ancestral proteins [29]. These ancestral proteins can then be characterized, providing evolutionary clues to sequence–function relationships in a growing number of protein systems, including classic models, e.g., recent work on the origin of cooperativity in hemoglobin [30]. A number of studies suggest that ancestral proteins are less selective for target ligands and more thermostable than extant sequences

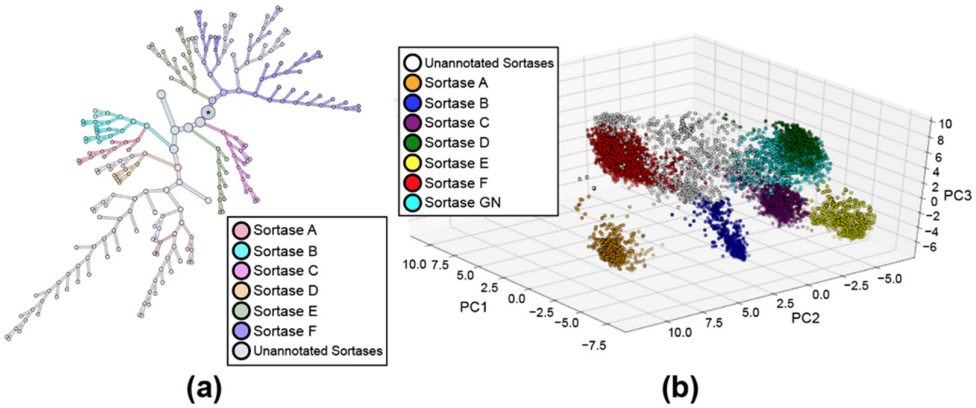
[31]. Therefore, we propose that ASR can be used as a method for identifying improved sortase sequences for protein engineering.

Here, we used principal component analysis (PCA) and ASR to study the sortase superfamily and Class A sortase sequence variation, respectively. Using PCA, we show that the main source of natural variation within sortase families occurs in a number of structurally-conserved loops near the active site. Using ASR, we characterized ancestral proteins of the genera *Staphylococus* and *Streptococcus*. While our ancestral *Staphylococcus* protein revealed lower relative activity than saSrtA, the ancestral *Streptococcus* enzyme had the second-highest activity of the four *Streptococcus* SrtA proteins studied in similar experiments [14,15]. Interestingly, the ancestral *Streptococcus* SrtA showed markedly increased activity and P1 promiscuity, as compared to its extant *S. pneumoniae* relative [14,15]. Although ancestral sortases from nodes that included multiple genera were expressed and purified, these enzymes were catalytically inactive, due to a number of potential factors. Overall, our work suggests that the ancestral *Streptococcus* protein was relatively more active as compared to its extant relatives and that the ASR technique provides a viable approach for exploring sequence variation in sortases from the same genera.

2. Results

2.1. Principal Component Analysis (PCA) of Bacterial Sortases

In order to gain a better understanding of global sequence patterns in the sortase superfamily, we used PCA to group and analyze 39,188 sortase sequences from all classes. This work builds off of recent studies that utilized a sequence similarity network to classify sortases [27]. Briefly, we downloaded all sequences annotated as "sortase" from UniProt and aligned them by MAFFT, followed by PCA [32,33]. The amino acids in each sequence were then classified by five parameters: hydrophobicity, disorder propensity, molecular weight, charge, and occupancy (defined as a binary value, where 1 = amino acid and 0 = insertion or deletion (indel) at this position) [34,35]. PCA was then performed on the resulting matrix. For visualization purposes this data was projected onto the first three principal components which describe 42.7% of the total variance (Figure S1a). Additionally, we performed Hierarchical Gaussian Mixture Model clustering of the sortase superfamily, as described in the Materials and Methods. On the entire principal component space we hierarchically fit a two Gaussian mixture model to the data until each subcluster reached a minimum size or the Gaussian mixture modeling process failed to identify two distinct Gaussians [36]. The resulting tree from this process can accurately distinguish the known sortase classes, as well as extract small subclusters of sortases and present them in a readable manner (Figure 1a). We also plotted our PCA using the top three principal components (Figure S1b). For visualization, we ran PCA on a subset of the data, including 9427 sequences that were filtered for low numbers of indels and manually verified (Figure 1b).



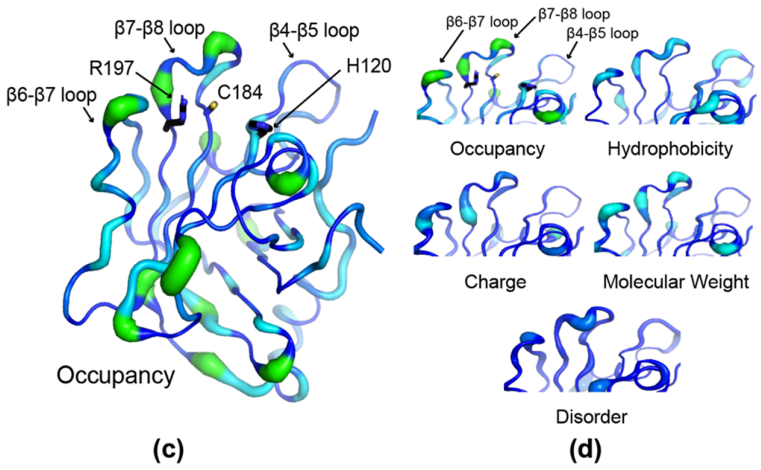


Figure 1. Principal component analysis (PCA) of sortase superfamily reveals sequence variability in structurally-conserved loops. (**a**) Hierarchical clustering of the sortase superfamily by Gaussian mixture model unsupervised classification on the PCA matrix distinguishes the known classes of sortases [36]. (**b**) Visualization of a subset of 9,427 of the high confidence sortase sequences plotted using principal components 1–3, to act as quality control of the data. The sequence is colored by which class of sortase it is annotated as by UniProt, when available. An equivalent plot of all 39,188 sortase sequences is in Figure S1d. (**c**,**d**) The five characteristics assigned numerical values in the PCA are visualized by width (from 0 to 1) and color (where dark blue is a value closer to 0 and green indicates a value closer to 1 using PyMOL. The *Streptococcus pyogenes* SrtA structure (PDB ID 3FN5) is used as the model to show variance in a "typical Class A sortase." The catalytic residues (H142, C208, and R216) are shown as side chain sticks and colored by heteroatom. The three structurally conserved loops that are discussed in this work are labeled. We focused on variance near the active site here, but notably, there is also a relatively large degree of variance on the other side of the protein (also Figure S1d).

This analysis verified previous classifications of sortases based on sequence alignment, network, and phylogenetic tree analyses [27,28,37]. For example, principal component 1 (PC1) separates the sortase F proteins from the rest of the superfamily and PC2 captures the separation between sortase B and the other sortase families, as well as sortase E and sortase A. These analyses allowed us to identify the regions of highest variability within each class based on the parameters defined above. We plotted our data onto previously determined sortase A structures by taking the distance from the centroid for each position in the multiple sequence alignment (Figure S1c). Consistent with expectations, we found that secondary structure elements are highly conserved, including the "sortase fold" β-barrel core and class-specific α-helices (Figures 1c–d and S1d). Additionally, PCA revealed that the highest degree of variability occurs in structurally conserved loops adjacent to the substrate recognition pocket (Figures 1c–d and S1d).

Given that the β6-β7 loop has been shown to be intimately involved in sortase substrate recognition in Staphylococcus aureus SrtA (saSrtA), we were intrigued that PCA revealed similar levels of variability in the β4–β5 and β7–β8 loops [38]. In the case of β7-β8, we were also motivated by previously reported mutations in the β7–β8 loop of saSrtA that have been shown to dramatically modulate sortase reaction rates [8,21,39]. Indeed, our work confirms that the β7–β8 loop dramatically affects the activity and substrate specificity of a sortase with narrow substrate tolerance, e.g., saSrtA, versus those that are more promiscuous, e.g., the Streptococcous SrtA proteins from S. pneumoniae, S. pyogenes, and S. agalactiae [14,15].

2.2. Ancestral Sequence Reconstruction of Class A Sortases

Building off our PCA analysis of sortase families, we wanted to further explore sequence space in these enzymes by performing ancestral sequence reconstruction (ASR) on Class A sortases. As detailed in the Materials and Methods, ultimately 400 sequences—including 7 SrtB sequences used to anchor the resulting phylogenetic tree—were used for the sequence reconstruction (Figure 2). Initially, we chose to characterize ancestral proteins at ancestral nodes for two SrtA genera with well-characterized family members, *Staphylococcus* and *Streptococcus* (Figure 2). We will refer to these proteins as ancStaphSrtA and ancStrepSrtA, respectively.

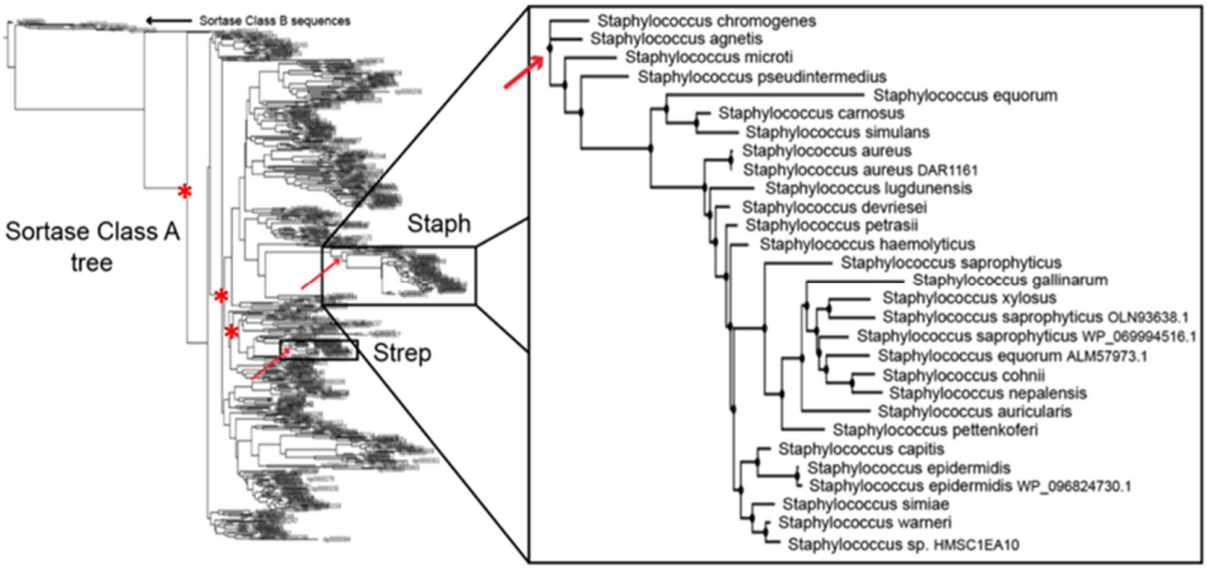


Figure 2. Constructed phylogenetic tree of SrtA sequences used for ancestral sequence reconstruction. The locations of ancStaphSrtA (also, inset) and ancStrepSrtA are indicated by red arrows. More ancestral proteins are indicated by asterisks; moving from left to right, these are ancNode-408, ancNode-503, and ancNode-547.

Multiple sequence alignments of our ancestral proteins with representative extant sequences reveals approximate values of 78.3% identity for ancStaphSrtA with another *Staphylococcus* SrtA sequence, and 69.5% identity for ancStrepSrtA with another *Streptococcus* SrtA sequence (Table 1, Figure 3). These are similar values to pairwise sequence alignment identities for most of the other representative extant sequences chosen (Table 1, Figure 3). These values represent an average of 30.67 and 49.67 mutations over the aligned regions for ancStaphSrtA and ancStrepSrtA with representative extant sequences, respectively.

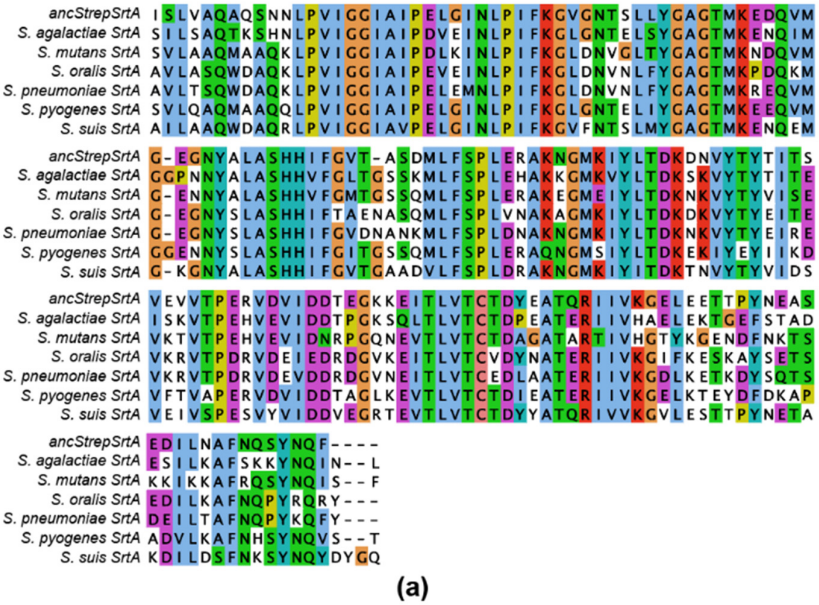
Table 1. Sequence identities of ancStaphSrtA and ancStrepSrtA with extant sequences. Number of identical residues out of total residues in the alignment are in the parentheses. (**a**) *Staphylococcus* SrtA sequence identities. (**b**) Streptococcus SrtA sequence identities.

(a)	***ancStaph***	***S. agnetis***	***S. aureus***	***S. auricularis***	***S. capitis***	***S. epidermidis***	***S. pettenkoferi***
ancStaph	X						
S. agnetis	68% (97/142)	X					
S. aureus	86% (127/147)	76% (108/142)	X				
S. auricularis	73% (107/146)	62% (91/146)	70% (101/144)	X			
S. capitis	88% (129/146)	62% (88/141)	84% (122/146)	68% (96/141)	X		

S. epidermidis	81% (116/143)	59% (84/142)	78% (114/146)	69% (98/142)	84% (124/147)	X	
S. pettenkoferi	74% (105/141)	58% (81/140)	67% (96/143)	78% (109/140)	73% (104/143)	71% (101/143)	X

(b)

	ancStrep	***S. agalactiae***	***S. mutans***	***S. oralis***	***S. pneumoniae***	***S. pyogenes***	***S. suis***
ancStrep	X						
S. agalactiae	68% (113/165)	X					
S. mutans	65% (106/163)	64% (109/169)	X				
S. oralis	69% (113/163)	58% (97/166)	65% (107/165)	X			
S. pneumoniae	68% (111/164)	57% (95/167)	64% (107/166)	81% (136/167)	X		
S. pyogenes	71% (117/164)	65% (109/168)	70% (117/168)	63% (104/166)	63% (105/166)	X	
S. suis	76% (125/164)	58% (98/168)	60% (101/168)	61% (101/165)	62% (103/166)	63% (104/166)	X



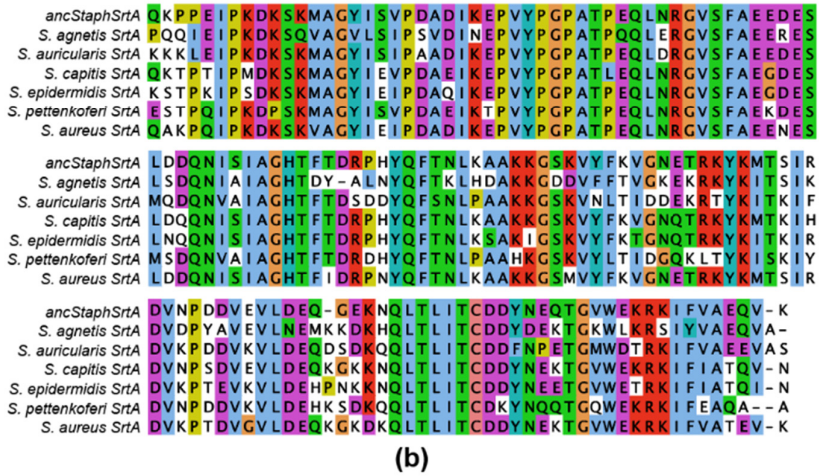


Figure 3. Multiple sequence alignments of ancestral proteins with representative (**a**) *Staphylococcus* and (**b**) *Streptococcus* SrtA proteins. Multiple sequence alignments of example *Staphylococcus* and *Streptococcus* SrtA proteins reveal a number of regions with high similarity.

To characterize the ancestral proteins, we recombinantly expressed and purified these enzymes and ran activity assays, as described in the Materials and Methods. Wild-type proteins were expressed and purified as previously described [14]. All protein sequences are included in the Supplemental Information. Briefly, activity was assessed using a FRET-based assay utilizing probes with a 2-aminobenzoyl fluorophore (Abz) and a 2,4-dinitrophenyl quencher (Dnp) on either side of a substrate motif, e.g., Abz-LPATGG-K(Dnp), where the P1′ Gly is in bold [8,14,15,26,40]. We varied the P1′ position to all amino acids and tested the two proteins (Figure 4). Similar data for spSrtA was previously reported [14]. Consistent with known target sequences and our previous work, our results revealed that although its activity was reduced >2-fold, ancStaphSrtA was similar to saSrtA and remained selective for a P1′ Gly residue [14,25,26]. In contrast, ancStrepSrtA could recognize several amino acids at the P1′ position, showing increased promiscuity, specifically Ala, Cys, Asn, and Ser, in addition to Gly [14,15,25,26]. As compared to spSrtA, the activity of ancStrepSrtA was increased ~2–3-fold. In comparison to our previous work on the in vitro activity of other *Streptococcus* SrtA proteins, ancStrepSrtA also appeared to be more active than *S. agalactiae* SrtA (sagSrtA), but less active than *S. pyogenes* SrtA (spySrtA) [15].

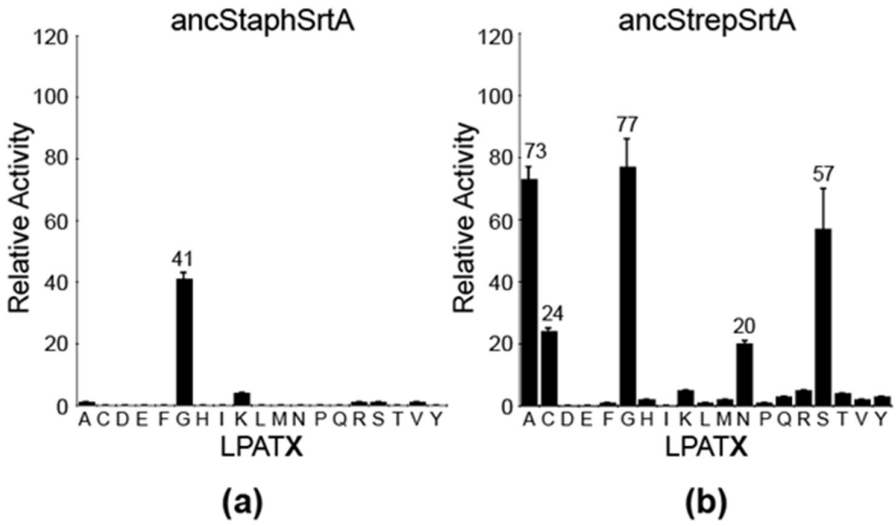


Figure 4. Relative enzyme activities for ancestral proteins. Substrate selectivity data for ancStaphSrtA (**a**) and ancStrepSrtA (**b**), proteins are shown as bar graphs. Substrate cleavage monitored via an increase in fluorescence at 420 nm from reactions of fluorophore-quencher probes with the generic structure Abz-LPATXG-K(Dnp) (LPAT**X**) in the presence of hydroxylamine. Bar graphs represent mean normalized fluorescence (± standard deviation) from at least three independent experiments, and average activity values over 10% relative activity are labeled. Assays are normalized against a previous set of reactions of saSrtA with LPAT**G** [14], to ensure consistency across data sets. Wild-type SaSrtA only recognizes LPAT**G**, as previously shown [25].

2.3. Structural Analyses of Ancestral SrtA Proteins

In order to better interpret our biochemical data, we used homology modeling to predict the structures of our ancestral proteins [41–43]. The template structure used for ancStaphSrtA was saSrtA-LPAT* (PDB ID 2KID), as this is the only known saSrtA structure in the active conformation, i.e., including Ca^{2+}, to our knowledge [44]. For ancStrepSrtA, we used a structure of spySrtA bound to the LPATS target sequence (PDB ID 74S0) as the template. Next, we compared our ancestral proteins to the wild-type SrtA enzymes used in our activity assays, highlighting variant residues (Figure 5a,b).

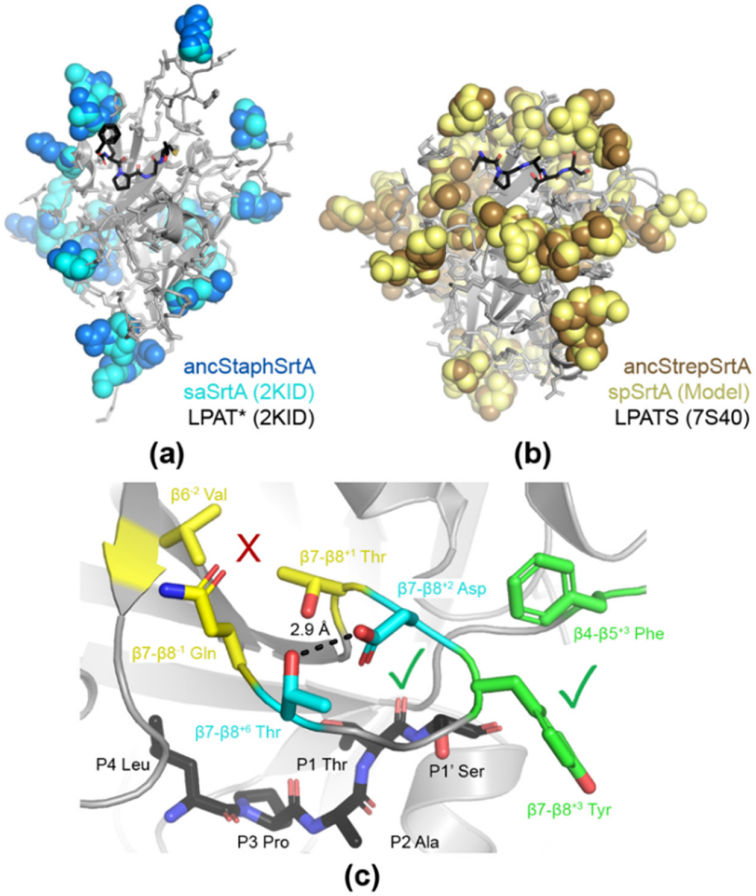


Figure 5. Structural comparison of ancestral protein models with wild-type SrtA proteins**.** The amino acid differences between wild-type saSrtA/ancStaphSrtA (**a**) and wild-type spSrtA/ancStrepSrtA (**b**) are represented as side chain spheres and colored as labeled. All four proteins are shown in gray cartoon representation with identical side chains as sticks. Ligands are shown in black sticks and colored by heteroatom (C = black, O = red, N = blue) and are the peptidomimetic LPAT* (**a**) and LPATS (**b**). Relevant PDB ID codes are in parentheses. (**c**) Interactions mediated by β7–β8 residues described in the text are highlighted by color, and the X and checkmarks are used to indicate whether they are predicted to be present. The ancStrepSrtA homology model is in gray cartoon, with the LPATS peptide (from PDB ID 7S40) rendered as in (**b**). Residues implicated in each interaction are shown as sticks and colored by heteroatom (1) β7–β8 loop with β6 position interaction (C = yellow), (2) β7–β8 intra-loop hydrogen bond (C = cyan), and (3) β7–β8 loop with β4–β5 position interaction (C = green).

Comparison of ancStaphSrtA to saSrtA (PDB ID 2KID) revealed very few changes near the active sites of these proteins (Figure 5a). The most concentrated amino acid variations in the vicinity of the LPAT* peptidomimetic are in the β6–β7 loop, with 6 differences amongst 16 residues. Notably, the saSrtA loop is 17 residues in length, so the ancStaphSrtA loop is truncated by one amino acid. This loop has previously been implicated in selectivity differences for saSrtA, suggesting that this loop variation may contribute to the over two-fold lower activity we see in ancStaphSrtA as compared to saSrtA (Figure 4) [38].

In our analysis of ancStrepSrtA and spSrtA, we used a previously generated homology model of spSrtA, as the only available structures are of a domain-swapped dimer whose activity has yet to be confirmed (Figure 5b) [14]. Notably, alignment of our homology model with the predicted structure from the AlphaFold Protein Structure Database reveals a root mean squared deviation (RMSD) for main chain atoms of 0.501 Å (489

atoms), with the largest amount of variation in the β6–β7 and β7–β8 loops (Figure S2) [45,46]. Although these sequences are less similar than the *Staphylococcus* proteins, we again observe relatively few amino acid substitutions in residues that directly interact with the ligand (Figure 5b). Here, we used the LPATS peptide from spySrtA-LPATS (PDB ID 7S40) for reference. We do observe amino acid variants in the β7–β8 loop residues of ancStrepSrtA as compared to spSrtA that may explain the increased activity of the ancestral protein. As we have previously described, an interaction between the $\beta6^{-2}$ (or two residues from the C-terminus of the β6 strand) R184 and two residues in the spSrtA β7-β8 loop, $\beta7–\beta8^{+1}$ (or 1 residue C-terminal to the catalytic Cys) E208 and $\beta7–\beta8^{-1}$ (or 1 residue N-terminal to the catalytic Arg) E214, weakens the overall activity of spSrtA [14]. In contrast, spySrtA does not contain this interaction and shows much higher relative activity [15]. AncStrepSrtA contains a $\beta7–\beta8^{+1}$ Thr, $\beta7–\beta8^{-1}$ Gln, and $\beta6^{-2}$ Val, suggesting that this interaction is also not present in this protein (Figure 5c). We do, however, observe that ancStrepSrtA likely conserves the two favorable interactions previously described that are mediated by β7–β8 loop residues, including an intra-loop hydrogen bond between $\beta7–\beta8^{+2}$ Asp and $\beta7–\beta8^{+6}$ Thr, as well as a hydrophobic interaction between the $\beta7–\beta8^{+3}$ Tyr residue and $\beta4–\beta5^{+3}$ Phe (or three residues C-terminal to the catalytic His) (Figure 5c) [14,15].

2.4. Investigating Ancestral Proteins at Distant Nodes

Finally, we wanted to test the activity of ancestral SrtA proteins at more distant nodes in our ASR analyses. We chose three sequences with relatively low sequence identity to ancStaphSrtA and ancStrepSrtA that were also distinct from each other (Figure 2, Table 2). All protein sequences are in the Supplemental Information. We named the proteins for their node characterization in the ASR, ancNode-408, ancNode-503, and ancNode-547.

Table 2. Comparative pairwise sequence identities of ancestral proteins in this study.

	ancStaph	***ancStrep***	***ancNode-408***	***ancNode-503***	***ancNode-547***
ancStaph	X				
ancStrep	30% (35/117)	X			
ancNode-408	35% (47/133)	51% (77/151)	X		
ancNode-503	33% (50/150)	56% (76/136)	78% (156/200)	X	
ancNode-547	54% (64/118)	59% (85/145)	64% (147/199)	77% (168/199)	X

We expressed and purified these proteins as described in the Materials and Methods. Notably, only fractions corresponding to the monomeric peak were retained following size exclusion chromatography, and based on their migration, these proteins are not aggregated and retain a similar radius of gyration as the wild-type proteins (Figure S3). Unfortunately, when evaluated using our FRET-based assay, all three proteins were catalytically inactive for sequences containing P1′ Ala, Gly, and Ser residues. Multiple sequence alignment of the ancestral proteins in this study suggests why these proteins may be catalytically inactive (Figure 6). Specifically, the manual refinement of the multiple sequence alignment used for ASR aimed to reduce numbers of gaps in the overall alignment, thereby optimizing alignments in areas of conserved secondary structure elements, e.g., the eight-stranded β-barrel structure conserved in the characterized sortase fold [9,14]. In doing so, we predict this introduced gaps in the structurally-conserved loops near the active site, e.g., the β4–β5 and β7–β8 loops previously mentioned here (Figure 6). The β6–β7 loop appears largely conserved in length, perhaps indicative of a higher degree of length conservation in this structural feature, as well as the β3–β4 loop, which, while spatially more distant from the active site, contains residues previously implicated in ligand recognition (Figure 6) [44].

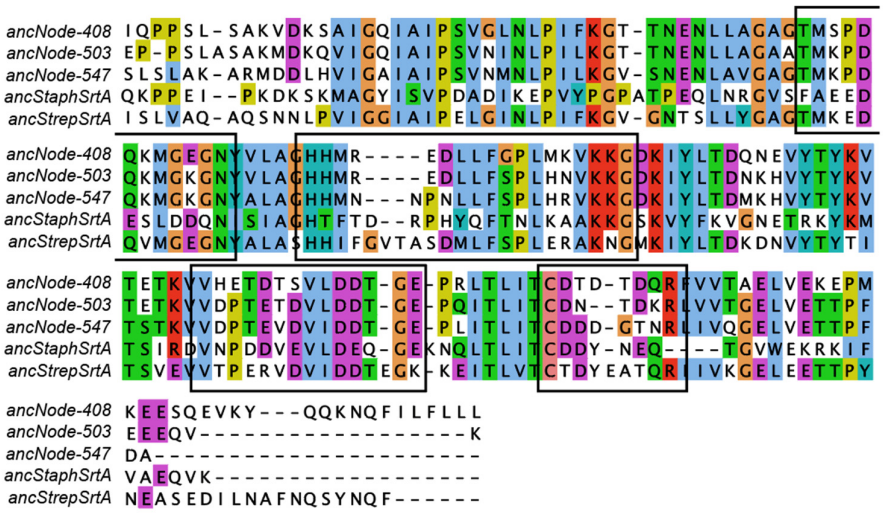


Figure 6. Multiple sequence alignment of ancestral proteins used in this study. Multiple sequence alignment of the ancestral proteins used in this study reveals the largest sequence variations within the β4–β5 and β6–β7 loops. The loop regions are indicated in black boxes, moving left-to-right and top-to-bottom; the order of these are β3–β5, β4–β5, β6–β7, and β7–β8 loops. The His at the first position of the β4–β5 loop, and the Cys and Arg at the first and last positions of the β7–β8 loop, respectively, are the catalytic residues. Notably, the β7–β8 loop of ancStaphSrtA does not align well with the other sequences, due to its increased length; this feature is also true of saSrtA with other SrtA proteins (data not shown).

The loop lengths of the β7–β8 loops of Class A sortases can vary quite dramatically. We previously characterized the β7–β8 loops of several Class A sortases which varied from 7 residues in *Streptococcus* proteins to 12 residues in *Staphylococcus aureus* [14]. These differences are also seen in the β4–β5 loop sequences, where saSrtA contains an additional three residues between the catalytic His residue and a conserved Phe residue as compared to spySrtA. These differences in lengths of critical loops likely hindered our ability to accurately reconstruct ancestral proteins at nodes that included multiple genera as descendants.

3. Discussion

In this work, we used bioinformatics tools, PCA and ASR, to investigate sequence variation in sortases. We used PCA to identify sources of variation in the sortase superfamily, recapitulating previous work that characterized the different classes based on distinct sequence properties [27,47]. In addition, we found that within each class, the largest variation typically occurs in structurally conserved loops, including those near the active site of the enzymes (Figure 1).

These loops were further implicated as sources of relatively high variation in our ASR biochemical studies. Here, while we were able to express and purify enzymatically active ancestral proteins of the *Staphylococcus* and *Streptococcus* genera, ancestral sequences of nodes that combined multiple genera were catalytically inactive. We predict that this is due to truncations in the β4–β5 and β6–β7 loops as a result of manual multiple sequence alignment refinement during ASR (Figure 6). Despite this, our ancStrepSrtA protein was the second-most active *Streptococcus* protein of four studied using similar activity assays [14,15]. We also find that ancStrepSrtA contains a similar degree of variation, ~30%, as other *Streptococcus* proteins with each other (Table 1b). For example, spySrtA shares 65% sequence identity with sagSrtA (109/168 residues), and 63% (105/166) with spSrtA. SpSrtA and sagSrtA are 57% identical (95/167). In contrast, spSrtA, spySrtA, and sagSrtA are 34% (34/121), 26% (38/148), and 35% (42/119) identical to saSrtA, respectively.

While ancestral proteins at deep nodes that included multiple genera as descendants were found to be inactive, the fact that they were able to be expressed and purified using the same methods as those used for extant proteins suggested that the central sortase fold remained intact. Future work to repeat the ASR with careful attention to loop lengths, as well as introduction of extant β4–β5 and β6–β7 loops into ancestral proteins, could provide a means for restoring activity to these enzymes, and may elucidate additional molecular characteristics of the contribution of these individual regions to the activity and selectivity of sortases. Such information would be very useful in future design efforts for sortase enzymes with improved catalytic efficiency or altered specificity. It would also be interesting to perform structural studies on these ancestral proteins, providing insights into potential differences compared to extant proteins with respect to the stereochemistry of target recognition.

There are a number of potential tools that can be used to examine sequence variation in bacterial sortases. Here, we utilized network and evolutionary approaches to investigate natural sequence variation. We argue that with the existence of thousands of sortase enzymes in multiple classes, there is still much to be discovered in extant sortase sequences [27,48]. In addition, directed evolution has proved to be an exciting technique to engineer sortase variation in vitro [17,21,49]. Both approaches, investigating natural sequences, as well as introducing new variation, will allow for a deeper understanding of the sequence determinants of activity and target selectivity, and can profoundly impact the study of the sortase enzyme family, both in protein engineering and for therapeutic uses.

4. Materials and Methods

Principal component analysis (PCA). Initial sequences were obtained from UniProt and an alignment was generated by MAFFT [32,33]. Initially, each sequence was given a score for the number of gaps present for each residue and the filtered alignment was realigned by MAFFT version 7. Subsequent analysis included all sequences without taking gaps into consideration (Figure 2b vs. Figures 2a and S2c). The sortase multiple sequence alignment (MSA) was converted to a tensor of sequences, characterized by MSA position and chemical property of each amino acid. Each amino acid was associated with 4 biochemical traits and a binary trait occupancy, as described. Each trait was normalized to the range from zero to one. In addition, gaps were given the average value of the matrix column with the exception of occupancy so that they would not contribute to variance of the column. Gapped positions were given an occupancy score of zero (for the other chemical properties gapped positions received the average score). After translating the MSA, the resulting tensor was flattened to matrix stacking of the chemical properties and was re-centered so that the matrix had a column-wise mean of zero. Principal component analysis was performed on the matrix by the singular value decomposition algorithm provided in the scikit learn Python package [50]. Clustering was performed by a Gaussian mixture model provided in the scikit-learn 1.1 Python package [50]. Optimal cluster numbers were scored by Bayesian information criterion. Visualization was performed using a script written in Python with matplotlib. Programs were run using default parameters, unless otherwise noted.

Ancestral Sequence Reconstruction. Nonredundant sortase sequences were sourced from the NCBI protein database [51]. Cluster Database at High Intensity with Tolerance program (CD-HIT) was used to filter out highly similar (>95%) identical sequences sourced from NCBI [52,53]. An all-vs-all basic local alignment search tool (BLAST) was used on the remaining sortase sequences, producing a sortase network which informed the assignment of sortase class groups (A–F) by using labeled sortase sequences to assign a class to each grouping [54]. Proteins surrounding the class A group were selected and an additional round of filtering was performed, where all highly similar proteins (>90%) were filtered out via CD-HIT. The remaining pool of sortase sequences was then subjected to alignment by MUltiple Sequence Comparison by Log-Expectation (MUSCLE), and then

manually curated to remove any outlying sequences [55]. Seven Class B sortase sequences (from *Streptococcus suis*, *Streptococcus oralis*, *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Bacillus anthracis*, *Listeria monocytogenes*, and *Enterococcus faecalis*) were added to anchor the resulting phylogenetic tree. The final alignment contained a total of 400 sequences. SrtA structures sourced from the PDB database were structurally aligned and sequence similarity between structural sequences (via PDB) and sortase sequences from the multi-sequence alignment (MSA) (via ASR) then informed the true alignment of the MSA. A phylogenetic tree was constructed from the MSA via phyml 3.0 and ancestral sequences were then generated at each node via multi-channel access XML (maxml) [56]. These latter steps were run using a python script. The aLRT values for proteins characterized were ancStaphSrtA = 15.6525, ancStrepSrtA = 13.0091, ancNode-408 = 17.7893, ancNode-503 = 28.8809, and ancNode-547 = 17.5286. Programs were run using default parameters, unless otherwise noted.

Protein expression and purification. Recombinant ancestral proteins (ancStaphSrtA, ancStrepSrtA, ancNode-408, ancNode-503, and ancNode-547) were expressed using *Escherichia coli* BL21 (DE3) cells in the pET28a(+) vector (Genscript), as previously described [14,15]. Transformed cells were grown at 37 °C in LB media to an OD_{600} 0.6–0.8, followed by induction using 0.15 mM IPTG for 18–20 h at 18 °C. The cells were harvested in lysis buffer [0.05 M Tris pH 7.5, 0.15 M NaCl, 0.5 mM ethylenediaminetetraacetic acid (EDTA)] and whole cell lysate was clarified using centrifugation, followed by filtration of the supernatant. The supernatant was initially purified using a 5 mL HisTrap HP column (Cytiva), with wash [0.05 M Tris pH 7.5, 0.15 M NaCl, 0.02 M imidazole, 0.001 M TCEP] and elution [wash buffer with 0.3 M imidazole] buffers.

Following immobilized metal affinity chromatography, the protein was concentrated using an Amicon Ultra-15 Centrifugal Filter Unit (10,000 NWML) followed by size exclusion chromatography (SEC) using a HiLoad 16/600 Superdex 75 column (Cytiva), with SEC running buffer [0.05 M Tris pH 7.5, 0.15 M NaCl, 0.001 M TCEP]. Purified protein fractions corresponding to the monomeric peak were pooled and concentrated. Purity was assessed using SDS-PAGE. Protein concentrations were determined using theoretical extinction coefficients calculated using ExPASy ProtParam [57]. Protein not immediately used was flash-frozen in SEC running buffer and stored at −80 °C.

Fluorescence Assay for Sortase Activity. Model peptide substrates with the general structure Abz-LPATXG-K(Dnp) (Abz = 2-aminobenzoyl, Dnp = 2,4-dinitrophenyl) were synthesized and purified as previously described [14]. Reactions were analyzed using a Biotek Synergy H1 plate reader as previously described [14,15]. Briefly, reactions were performed a 100 μL reaction volume consisting of 5 μM sortase, 50 μM peptide substrate, 5 mM hydroxylamine nucleophile, and 10% (*v*/*v*) 10x sortase reaction buffer (500 mM Tris pH 7.5, 1500 mM NaCl, and 100 mM $CaCl_2$). The reactions were performed in triplicate and the fluorescence intensity of each well was measured at 2-min time intervals over a 2-hour period at room temperature (λ_{ex} = 320 nm, λ_{em} = 420 nm, and detector gain = 75). For each substrate sequence, the background fluorescence of the intact peptide in the absence of enzyme was subtracted from the observed experimental data. Background-corrected fluorescence data was then normalized to the fluorescence intensity of a benchmark reaction between wild-type saSrtA and Abz-LPATGG-K(Dnp), as previously described [14,15].

Structural analyses. Alignments were visualized using AliView [58]. Phylogenetic trees were visualized with FigTree v1.4.3 [59]. Homology modeling was performed using the SwissModel web interface [41,43]. Structural analyses and figure rendering were done using PyMOL. Enzyme assay graphs were prepared using Kaleidagraph. The *Streptococcus pneumoniae* SrtA structure was downloaded from the AlphaFold Protein Structure Database (entry number Q8DPM3) [45,46].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bacteria1020011/s1, Figure S1: Principal Component Analysis (PCA) of sortase superfamily; Figure S2: Structures of spSrtA from the AlphaFold database and homology modeling; Figure S3: Size exclusion chromatography of ancestral SrtA proteins; Recombinant protein sequences used in this study [14,25].

Author Contributions: Conceptualization, J.D.V., Z.R.S., M.J.H. and J.F.A.; methodology, J.D.V., Z.R.S., M.J.H. and J.F.A.; software, J.D.V., Z.R.S., and M.J.H.; validation, S.A.S., I.M.P., J.E.S., D.A.J. and B.A.V.; formal analysis, J.D.V., J.M.A. and J.F.A.; investigation, J.D.V., S.A.S., Z.R.S., I.M.P., J.E.S., D.A.J. and B.A.V.; resources, J.M.A., M.J.H. and J.F.A.; data curation, J.D.V. and J.F.A.; writing—original draft preparation, J.F.A.; writing—review and editing, J.D.V., S.A.S., I.M.P., J.E.S., J.M.A., M.J.H. and J.F.A.; visualization, J.D.V. and J.F.A.; supervision, J.M.A., M.J.H. and J.F.A.; project administration, J.F.A.; funding acquisition, J.M.A., M.J.H. and J.F.A. All authors have read and agreed to the published version of the manuscript.

Funding: J.F.A. and J.M.A. were both funded by Cottrell Scholar Awards from the Research Corporation for Science Advancement. J.F.A. was also funded by NSF CHE-CAREER-2044958. M.J.H. was funded by NSF DEB-CAREER-1844963. In addition, I.M.P. received an Elwha Undergraduate Summer Research Award and D.A.J. received a Joseph & Karen Morse Student Research in Chemistry Fellowship to fund summer research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sizar, O.; Unakal, C.G. Gram Positive Bacteria. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
2. Diekema, D.J.; Hsueh, P.-R.; Mendes, R.E.; Pfaller, M.A.; Rolston, K.V.; Sader, H.S.; Jones, R.N. The Microbiology of Bloodstream Infection: 20-Year Trends from the SENTRY Antimicrobial Surveillance Program. *Antimicrob. Agents Chemother.* **2019**, *63*, e00355-19.
3. Maharath, A.; Ahmed, M.S. Bacterial Etiology of Bloodstream Infections and Antimicrobial Resistance Patterns from a Tertiary Care Hospital in Malé, Maldives. *Int. J. Microbiol.* **2021**, *2021*, 3088202.
4. Zhu, Q.; Yue, Y.; Zhu, L.; Cui, J.; Zhu, M.; Chen, L.; Yang, Z.; Liang, Z. Epidemiology and microbiology of Gram-positive bloodstream infections in a tertiary-care hospital in Beijing, China: A 6-year retrospective study. *Antimicrob. Resist. Infect. Control* **2018**, *7*, 107.
5. Vollmer, W.; Blanot, D.; de Pedro, M.A. Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* **2008**, *32*, 149–167.
6. Feng, D.F.; Cho, G.; Doolittle, R.F. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 13028–13033.
7. Marraffini, L.A.; DeDent, A.C.; Schneewind, O. Sortases and the Art of Anchoring Proteins to the Envelopes of Gram-Positive Bacteria. *Microbiol. Mol. Biol. Rev.* **2006**, *70*, 192–221.
8. Ton-That, H.; Mazmanian, S.K.; Alksne, L.; Schneewind, O. Anchoring of surface proteins to the cell wall of Staphylococcus aureus. Cysteine 184 and histidine 120 of sortase form a thiolate-imidazolium ion pair for catalysis. *J. Biol. Chem.* **2002**, *277*, 7447–7452.
9. Jacobitz, A.W.; Kattke, M.D.; Wereszczynski, J.; Clubb, R.T. Sortase transpeptidases: Structural biology and catalytic mechanism. *Adv. Protein Chem. Struct. Biol.* **2017**, *109*, 223–264.
10. Spirig, T.; Weiner, E.M.; Clubb, R.T. Sortase enzymes in Gram-positive bacteria. *Mol. Microbiol.* **2011**, *82*, 1044–1059.
11. Antos, J.M.; Truttmann, M.C.; Ploegh, H.L. Recent advances in sortase-catalyzed ligation methodology. *Curr. Opin. Struct. Biol.* **2016**, *38*, 111–118.
12. Zhang, J.; Liu, H.; Zhu, K.; Gong, S.; Dramsi, S.; Wang, Y.-T.; Li, J.; Chen, F.; Zhang, R.; Zhou, L.; et al. Antiinfective therapy with a small molecule inhibitor of Staphylococcus aureus sortase. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 13517–13522.
13. Ton-That, H.; Liu, G.; Mazmanian, S.K.; Faull, K.F.; Schneewind, O. Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 12424–12429.
14. Piper, I.M.; Struyvenberg, S.A.; Valgardson, J.D.; Johnson, D.A.; Gao, M.; Johnston, K.; Svendsen, J.E.; Kodama, H.M.; Hvorecny, K.L.; Antos, J.M.; et al. Sequence variation in the β7–β8 loop of bacterial class A sortase enzymes alters substrate selectivity. *J. Biol. Chem.* **2021**, *297*, 100981.
15. Gao, M.; Johnson, D.A.; Piper, I.M.; Kodama, H.M.; Svendsen, J.E.; Tahti, E.; Longshore-Neate, F.; Vogel, B.; Antos, J.M.; Amacher, J.F. Structural and biochemical analyses of selectivity determinants in chimeric Streptococcus Class A sortase enzymes. *Protein Sci.* **2022**, *31*, 701–715.

16. Dai, X.; Böker, A.; Glebe, U. Broadening the scope of sortagging. *RSC Adv.* **2019**, *9*, 4700–4721.
17. Podracky, C.J.; An, C.; DeSousa, A.; Dorr, B.M.; Walsh, D.M.; Liu, D.R. Laboratory evolution of a sortase enzyme that modifies amyloid-β protein. *Nat. Chem. Biol.* **2021**, *17*, 317–325.
18. Bierlmeier, J.; Álvaro-Benito, M.; Scheffler, M.; Sturm, K.; Rehkopf, L.; Freund, C.; Schwarzer, D. Sortase-Mediated Multi-Fragment Assemblies by Ligation Site Switching. *Angew. Chem. Int. Ed.* **2022**, *61*, e202109032.
19. Kruger, R.G.; Otvos, B.; Frankel, B.A.; Bentley, M.; Dostal, P.; McCafferty, D.G. Analysis of the substrate specificity of the Staphylococcus aureus sortase transpeptidase SrtA. *Biochemistry* **2004**, *43*, 1541–1551.
20. Freund, C.; Schwarzer, D. Engineered sortases in peptide and protein chemistry. *Chembiochem* **2021**, *22*, 1347–1356.
21. Chen, I.; Dorr, B.M.; Liu, D.R. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 11399–11404.
22. Hirakawa, H.; Ishikawa, S.; Nagamune, T. Design of Ca^{2+}-independent Staphylococcus aureus sortase A mutants. *Biotechnol. Bioeng.* **2012**, *109*, 2955–2961.
23. Wójcik, M.; Vázquez Torres, S.; Quax, W.J.; Boersma, Y.L. Sortase mutants with improved protein thermostability and enzymatic activity obtained by consensus design. *Protein Eng. Des. Sel.* **2019**, *32*, 555–564.
24. Wójcik, M.; Szala, K.; van Merkerk, R.; Quax, W.J.; Boersma, Y.L. Engineering the specificity of Streptococcus pyogenes sortase A by loop grafting. *Proteins* **2020**, *88*, 1394–1400.
25. Nikghalb, K.D.; Horvath, N.M.; Prelesnik, J.L.; Banks, O.G.B.; Filipov, P.A.; Row, R.D.; Roark, T.J.; Antos, J.M. Expanding the Scope of Sortase-Mediated Ligations by Using Sortase Homologues. *Chembiochem* **2018**, *19*, 185–195.
26. Schmohl, L.; Bierlmeier, J.; von Kügelgen, N.; Kurz, L.; Reis, P.; Barthels, F.; Mach, P.; Schutkowski, M.; Freund, C.; Schwarzer, D. Identification of sortase substrates by specificity profiling. *Bioorg. Med. Chem.* **2017**, *25*, 5002–5007.
27. Malik, A.; Kim, S.B. A comprehensive in silico analysis of sortase superfamily. *J. Microbiol.* **2019**, *57*, 431–443.
28. Di Girolamo, S.; Puorger, C.; Castiglione, M.; Vogel, M.; Gébleux, R.; Briendl, M.; Hell, T.; Beerli, R.R.; Grawunder, U.; Lipps, G. Characterization of the housekeeping sortase from the human pathogen Propionibacterium acnes: First investigation of a class F sortase. *Biochem. J.* **2019**, *476*, 665–682.
29. Harms, M.J.; Thornton, J.W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **2010**, *20*, 360–366.
30. Pillai, A.S.; Chandler, S.A.; Liu, Y.; Signore, A.V.; Cortez-Romero, C.R.; Benesch, J.L.P.; Laganowsky, A.; Storz, J.F.; Hochberg, G.K.A.; Thornton, J.W. Origin of complexity in haemoglobin evolution. *Nature* **2020**, *581*, 480–485.
31. Wheeler, L.C.; Lim, S.A.; Marqusee, S.; Harms, M.J. The thermostability and specificity of ancient proteins. *Curr. Opin. Struct. Biol.* **2016**, *38*, 37–43.
32. UniProt Consortium UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515.
33. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* **2017**, *20*, 1160–1166.
34. Campen, A.; Williams, R.M.; Brown, C.J.; Meng, J.; Uversky, V.N.; Dunker, A.K. TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **2008**, *15*, 956–963.
35. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.
36. Schwartz, G.W.; Zhou, Y.; Petrovic, J.; Fasolino, M.; Xu, L.; Shaffer, S.M.; Pear, W.S.; Vahedi, G.; Faryabi, R.B. TooManyCells identifies and visualizes relationships of single-cell clades. *Nat. Methods* **2020**, *17*, 405–413.
37. Kattke, M.D.; Chan, A.H.; Duong, A.; Sexton, D.L.; Sawaya, M.R.; Cascio, D.; Elliot, M.A.; Clubb, R.T. Crystal Structure of the Streptomyces coelicolor Sortase E1 Transpeptidase Provides Insight into the Binding Mode of the Novel Class E Sorting Signal. *PLoS ONE* **2016**, *11*, e0167763.
38. Bentley, M.L.; Gaweska, H.; Kielec, J.M.; McCafferty, D.G. Engineering the substrate specificity of Staphylococcus aureus Sortase A. The beta6/beta7 loop from SrtB confers NPQTN recognition to SrtA. *J. Biol. Chem.* **2007**, *282*, 6571–6581.
39. Zou, Z.; Nöth, M.; Jakob, F.; Schwaneberg, U. Designed Streptococcus pyogenes Sortase A Accepts Branched Amines as Nucleophiles in Sortagging. *Bioconjug. Chem.* **2020**, *31*, 2476–2481.
40. Kruger, R.G.; Dostal, P.; McCafferty, D.G. Development of a high-performance liquid chromatography assay and revision of kinetic parameters for the Staphylococcus aureus sortase transpeptidase SrtA. *Anal. Biochem.* **2004**, *326*, 42–48.
41. Bordoli, L.; Kiefer, F.; Arnold, K.; Benkert, P.; Battey, J.; Schwede, T. Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* **2009**, *4*, 1–13.
42. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; de Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303.
43. Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics* **2006**, *22*, 195–201.
44. Suree, N.; Liew, C.K.; Villareal, V.A.; Thieu, W.; Fadeev, E.A.; Clemens, J.J.; Jung, M.E.; Clubb, R.T. The structure of the Staphylococcus aureus sortase-substrate complex reveals how the universally conserved LPXTG sorting signal is recognized. *J. Biol. Chem.* **2009**, *284*, 24465–24477.
45. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444.

46. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
47. Bradshaw, W.J.; Davies, A.H.; Chambers, C.J.; Roberts, A.K.; Shone, C.C.; Acharya, K.R. Molecular features of the sortase enzyme family. *FEBS J.* **2015**, *282*, 2097–2114.
48. Comfort, D.; Clubb, R.T. A comparative genome analysis identifies distinct sorting pathways in gram-positive bacteria. *Infect. Immun.* **2004**, *72*, 2710–2722.
49. Dorr, B.M.; Ham, H.O.; An, C.; Chaikof, E.L.; Liu, D.R. Reprogramming the specificity of sortase enzymes. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 13343–13348.
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830.
51. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2017**, *45*, D12–D17.
52. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659.
53. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152.
54. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
55. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
56. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321.
57. Wilkins, M.R.; Gasteiger, E.; Bairoch, A.; Sanchez, J.C.; Williams, K.L.; Appel, R.D.; Hochstrasser, D.F. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **1999**, *112*, 531–552.
58. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **2014**, *30*, 3276–3278.
59. FigTree. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed 16 March 2022).